ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/220046767

Regular Polysemy in WordNet

Article · January 2009

Source: DBLP

CITATIONS
10

READS
178

2 authors, including:

Francois-Regis Chaumartin
Paris Diderot University
16 PUBLICATIONS **222** CITATIONS

SEE PROFILE

All content following this page was uploaded by Francois-Regis Chaumartin on 20 May 2014.

The user has requested enhancement of the downloaded file.

Volume 24 - Number 2 - 2009 - ISSN 0175-1336

JLCL

Journal for Language Technology
and Computational Linguistics

Lexical-Semantic and Ontological Resources - Maintenance, Representation and Standards

Herausgegeben von / Edited by
C. Kunze, L. Lemnitzer, R. Osswald

GSCL Gesellschaft für Sprachtechnologie & Computerlinguistik

Contents

Editorial

This volume contains selected and revised papers presented at the workshop "Lexical-Semantic and Ontological Resources - Maintenance, Representation, and Standards" held in conjunction with KONVENS 2008 in Berlin on October 1st, 2008. The workshop was a follow-up of a series of thematically related events, starting with two GLDV workshops on GermaNet in 2003 and 2005 and continued by meetings on lexical-semantic resources at the DGfS 2006 and GLDV 2007 conferences. We aimed at providing a forum for discussing recent developments in the area of lexical-semantic resources - especially with regard to wordnets and ontologies. The focus was on the maintenance, extension, representation, and standardization of such resources.
We received 11 contributions of which 9 have been selected for presentation during the workshop. All extended abstracts submitted were reviewed by three members of the Program Committee. In a second round of reviewing, six contributions have been selected for publication in this special issue of the Journal for Language Technology and Computational Linguistics. The papers presented at the workshop have been grouped under three main thematic focuses:

1. Lexical semantics and extension of lexicons with new relations or features
2. Methods and tools for building and maintaining lexical resources
3. Representation issues and interoperability between different kinds of semantic resources

In each session three papers have been presented during the workshop of which two respectively are included in this volume.

The first paper by Lucie Barque and Francois-Regis Chaumartin deals with "Regular Polysemy in WordNet", proposing a method to extract regular polysemy patterns from Wordnet data in order to represent them in a broad coverage computational lexicon. This crucial issue has both been tackled in theoretical approaches to lexical semantics like that of Apresjan (1974) and formulated as a need in natural language processing (see Ravin and Leacock (2000)). After defining the basic notion of regular polysemy and giving an account of related work, the authors describe their methodology of creating polysemy patterns by exploiting the synset hierarchies and glosses of WordNet and present the results of an evaluation of their results. They detected more than 60 polysemy patterns, differentiating specialization, metaphor and metonymy.

Nofiza Vokhidova proposes a new lexical-semantic paradigmatic relation to be included into electronically available lexical resources such as elexiko. Her paper "Überlegungen zur Erweiterung lexikalisch-semantischer Ressourcen durch die Graduonymie" empirically describes graduality phenomena among members of word groups containing a feature to a lower or higher degree. These words could be ordered on a scale, as the author's example of "breeze", "wind", "storm" indicates. The paper assesses how graduonymy is captured in

lexical-semantic resources and how corpus-based studies and empirical data from query experiments can help extending computational lexicons. Encoding more fine-grained sense relations like graduonymy serves to more precisely and adequately capture structure of the lexical system of
a given language.

Kiril Simov's contribution "Ontology-Based Lexicon of Bulgarian" describes the construction of an ontology-driven lexicon, focusing on the Bulgarian language for which high quality semantic resources are still lacking. The lexicon is related to an upper ontology and domain ontologies by several mapping and alignment processes connecting grammatical, textual and ontological layers. For automatic word sense disambiguation, corpora are annotated with ontological information in order to train machine learning components. An annotation grammar is being developed for accurate matches between ambiguous textual words and concepts from the ontology. The work presented in the paper is a contribution towards a richer resource infrastructure for Bulgarian.

The paper of Ernesto William de Luca and Andreas Nürnberger titled "LexiRes RDF/OWL Editor: Maintaining Multilingual Resources" presents an RDF/OWL tool for editing and structuring ontologies with a focus on the use of multilingual resources such as the RDF/OWL version of EuroWordNet. The tool allows the user to merge similar concepts in and across languages as well as the multilingual search of synsets. Furthermore, wordnets can be enriched with external OWL resources by using LexiRes. The tool also includes a visualization component.

In their paper "Use, Re-Use and Synergetic Benefit: The interplay between WordNet and dicitionary data", Sanni Nimb and Lars Trap-Jensen emphasize the interplay between Danish Wordnet and Den Danske Ordbog (DDO). The Danish WordNet has been compiled on the basis of DDO definitions, but was also extended with further features and relations. The Danish Wordnet is used to support the search engine of DDO. The paper nicely demonstrates how different types of resources can be mutually exploited in order to enhance one another.

The final paper "Representing a Resource of Formal Lexical-semantic descriptions in the Web Ontology Language" by Fabienne Martin, Dennis Spohr and Achim Stein combines theoretical analyses of verb meanings with a formal representation (OWL) in order to allow for disambiguating verbs in context by logical reasoning. To this end, the selectional restrictions of verbs are determined based on mappings from the French EuroWordNet to SUMO and DOLCE. The Semantic Rule Language SWRL is then applied for calculating inferences.

The presentations of the workshop ignited lively discussions by the participants. Several lines and issues for further research were identified and the interest was expressed by several participants that these lines could be followed up and presented in successive workshops.

We would like to thank the Organizing Committee of KONVENS 2008 for making this workshop become possible and the Berlin-Brandenburgische Akademie der Wissenschaften for hosting the event in its beautiful building in charming mid-Berlin. We would also like to express our gratitude to Alexander Mehler and Christian Wolff for publishing the workshop proceedings as a special issue of the JLCL.

We would also like to thank the audience for contributing to the success of this workshop. Furthermore, we would like to thank the Program Committee for their reviews:

Jörg Asmussen (Kopenhagen)
Paul Buitelaar (Saarbrücken)
Christiane Fellbaum (Berlin, Princeton)
Piklu Gupta (Tübingen)
Marc Kemp-Snijders (Nijmegen)
Maciej Piasecki (Wroclaw)
Uwe Quasthoff (Leipzig)

Last but not least, we want to express our thanks to all speakers who presented their research on our workshop and the authors for compiling paper versions of their talks.

Claudia Kunze, Lothar Lemnitzer, Rainer Osswald

Lucie Barque, François Régis Chaumartin

Regular polysemy in WordNet

1 Introduction

The importance of describing regular polysemy in a lexicon has often been outlined, especially in the field of natural language processing (for a good overview of this issue, see (Ravin and Leacock, 2000)). Unfortunately, no existing broad-coverage semantic lexicon has been built following this relatively recent advice. And since producing a broad coverage semantic lexicon is a very time-consuming task, one has to put this idea into practice on existing lexicons. WordNet is an appropriate lexical semantic ressource for running this experiment as it is machine readable and has a wide coverage (Fellbaum, 1998). In this paper, we introduce a method to create regular polysemy patterns from WordNet data and to automatically detect their occurrences in the lexicon.

WordNet is structured as a hierarchy of synsets (sets of lexical units that are synonyms), representing lexical concepts. These concepts are linked by several kinds of semantic relations (hyperonymy, meronymy, antonymy, etc.). Each synset is also described by a lexicographic definition, as illustrated below (we denote by L#i the i^{th} lexical unit of a polysemic word in WordNet) :

- {treachery#2, betrayal#1, treason#3, perfidy#2} = *an act of deliberate betrayal*

As one can see in this example, the definition includes a word (*betrayal*) of the same form as one of the lexical units of the synset (betrayal#1). From now on, we will systematically distinguish the notion of **lexical unit** (one form associated to one meaning) from the notion of **word** (one form associated to possibly several meanings, in other terms, a polysemic unit). So we will say that the lexical unit `betrayal#1` is defined by another lexical unit of the word BETRAYAL but we don't know which one until the words of the definition are disambiguated[1]. In fact, WordNet contains implicit relations between pairs of lexical units that share the same form. Our goal is to make these relations explicit, and, when it is relevant, to generalize them in order to account for regular polysemy relations.

This paper consists of three sections. The first section defines the notion of regular polysemy and gives a quick overview of studies devoted to the description of regular polysemy in WordNet. The second one exposes our goals and explains our method to reach them, which consists in the "computer-assisted" creation of polysemy patterns.

[1] One has to mention the "eXtended WordNet" project, developed in 2003 at the University of Dallas (http://xwn.hlt.utdallas.edu/index.html). This project enhances WordNet 2.0 with a logical representation and a syntactic analysis of the definitions associated to the synsets. After the syntactic analysis, words from the definitions are disambiguated, with a *gold*, *silver* and *normal* quality (*gold* quality representing a manual evaluation). Since only 14 715 words out of 631 684 open-class words are gold quality (2,33%), we decided not to use this resource in our study.

The last section presents our results, which take the form of a classification of these patterns and a measure of their regularity in the most recent version of WordNet.

2 Regular polysemy in WordNet : an overview

In this section, we will first define the important notions our study is based on, mainly the notion of regular polysemy. We will then explain why it is important to describe this phenomenon in the lexicon. Finally, we will give a quick overview of studies that have been devoted to the description of polysemy in WordNet.

2.1 A definition of regular polysemy

We follow J. Apresjan's definition of regular polysemy :

> " Polysemy of a word A with the meaning a_i and a_j is called regular if, in the given language, there exists at least one other word B with the meaning b_i and b_j, which are semantically distinguished from each other in exactly the same way as a_i and a_j and if a_i and b_i, a_j and b_j are non-synonymous." (Apresjan, 1974):16

For instance, polysemy of the word CHERRY, with the meaning `fruit` and `color`, is regular since there exists at least one other word, CHESTNUT, that also has the meaning `fruit`[2] and `color`, as illustrated below with data from WordNet:

- {cerise#1, **cherry**#4} = *the red color of* ***cherries***
- {**chestnut**#4} = *the brown color of* ***chestnuts***

J. Apresjan gives us another important definition of a particular type of regular polysemy, that he calls **productive** and that we will call **systematic**, following the authors who have addressed this issue (Pustejovsky, 1995), (Buitelaar, 1998), among others.

> We will call a given type 'A' – 'B' of regular polysemy *productive*, if for any word which has the meaning 'A' it is true that it can be used also in the meaning 'B' (if 'A', then 'B') [...] Consequently, productivity is determined only by totality of scope of the units with the given combination of properties; the class itself of such units may be very small. (Apresjan (1992): 214)

The regular polysemy illustrated above between a lexical unit that denotes a `color` and a lexical unit that denotes a `fruit` is not systematic. Indeed, the example given under (1c) shows that it is difficult to generate a lexical unit that denotes a `color` from every lexical unit that denotes a `fruit`.

[2] Note that this data doesn't indicate that the word used in the definition denotes a `fruit`, since "*cherries*" and "*chestnuts*" are not disambiguated.

(1) a. *I like your **cherry** shirt.*

b. *I like your **banana** shirt.*

c. ? *I like your **pear** shirt.*

On the other hand, the regular polysemy between a lexical unit that denotes a `dish` and a lexical unit that denotes the `quantity` of food contained in this dish is a systematic one since you can derive the `quantity` (of food) sense from every word that denotes a `dish` (*a **bowl** of rice, a **spoon** of flour*, etc).

Between the minimal condition of regularity (at least two words that present the same sense alternation) and the systematicity (every word that denotes an `X` can also denote an `Y`), there is a large scale of regularities that a lexicon should be able to take into account.

It is one thing to consider the type of regularity of a given polysemy relation. Another thing is to consider the nature of this polysemy relation. In our study, we will consider the three following well-known categories of regular polysemy :

1. **Specialization** : a lexical unit L2 is a specialization of a lexical unit L1 if its meaning is more specific than that of L1. In the example given below, the lexical unit `pressure#7` denotes a particular kind of the pressure denoted in the definition.
   - {**pressure**#7} = *the **pressure** exerted by the atmosphere*
2. **Metaphor** : two lexical units L1 and L2 are in a metaphoric relation if L1 and L2's referents are in a relation of analogy, that is, if they are similar in almost one aspect (Johansen and Larsen, 2002). For example, the laugh denoted by `cackle#3` resembles a *hen's cackle*, as stated in the definition.
   - {**cackle**#3} = *a loud laugh suggestive of a hen's **cackle***
3. **Metonymy** : two lexical units L1 and L2 are in a metonymic relation if L1 and L2's referents are in a relation of contiguity, in other words if the two referents are "in contact", in the concrete or in the abstract sense of the word "contact". For example the relation between the two meanings of chestnut (`color` and `fruit`) is a metonymic relation since the color denoted by `chestnut#4` is the color of the fruit denoted by CHESTNUT in the definition.
   - {**chestnut**#4} = *the brown color of **chestnut***

2.2 Description of regular polysemy

The description of regular polysemy, by means of lexical rules (Ostler and Atkins, 1991), (Copestake and Briscoe, 1995) or by means of generative mechanisms during composition (Pustejovsky, 1995), presents at least two benefits.

From a theoretical point of view, the description of regular polysemy accounts for lexical creativity (Pustejovsky, 1995). For example, a lexicon that contains the description of the metaphoric regular polysemy relation between a lexical unit that denotes the fact of `tying an animal/person's part of the body` and a lexical unit that denotes the fact of `preventing someone from doing something` (regular polysemy that concerns TO MUZZLE, TO HOBBLE, TO GAG, etc) potentially contains the description of a new meaning. For example, the meaning of HANDCUFF, illustrated in (2)[3] below, is not listed in the consulted dictionaries (WordNet, Longman and Cobuild) and can thus be considered as a lexical creation. Having the description of the rule that has generated this meaning gives us information about this new use of HANDCUFF.

(2) *How can we build a 'Knowledge economy' if research* ***is handcuffed*** *?*

The second benefit, a more practical one, concerns a better methodology to develop the lexicon on which is based the description of regular polysemy (in our case, WordNet). Indeed, the description of regular polysemy by means of lexical rules allows for a more systematic encoding of the lexical data since they provide an underspecified definition to the lexicographer (Ostler and Atkins, 1991). For example, the underspecified definition, illustrated below in bold letters, can be used to define other lexical units of type `quantity` (*of food contained in a dish*), as illustrated below :

- L2 of X = **quantity of X contained in** L1
- plate#2 of X = **quantity of X contained in** plate#1
- bowl#2 of X = **quantity of X contained in** bowl#1

2.3 Description of regular polysemy in WordNet

Several studies have already been devoted to the description of regular polysemy in WordNet. We present some of them here, explaining how they converge and how they differ from our study.

Corelex is an ontology of 126 semantic types, each of them representing a systematic polysemous class of nouns (Buitelaar, 1998). This ontology aims to account for underspecification in discourse analysis, the author assuming that underspecification is often due to systematic polysemy. The method that has been employed to build this ontology is the following : the first step consists in delimitating a set of 39 basic types from the top level synsets of the noun hierarchy (for example, `entity`, `human`, `animal`, `food`, `act`, `event`, `location`, etc.). In the second step, each lexical unit from a polysemic word is assigned one of these basic types. For example, the seven senses of BOOK can be reduced to two basic types `artifact` and `communication`. The third step consists in grouping all polysemic words that share the same

[3] This sentence has been found on the web.

alternation of basic types. For example, CATALOGUE, HOROSCOPE, PRESCRIPTION, etc. share with BOOK the '`artifact-communication`' alternation. This step gives rise to 1648 classes, reduced to 529 polysemous classes once classes that contain only one member are removed. This 529 polysemous classes are then grouped into 126 semantic types. For example, the semantic type `cae` puts together the following polysemous classes : "`act∼artifact∼communication`" (CHOREOGRAPHY, DECORATION, DEVICE), "`act∼artifact∼communication∼psychological_feature`" (CALL, CONSTRUCTION, PORTRAYAL), "`act∼artifact∼psychological_feature`" (ARCHITECTURE, DESIGN, HABIT, ...), "`artifact∼communication`" (BOOK, CATALOGUE, HOROSCOPE), etc.

The method is interesting for at least two reasons. Firstly the Corelex approach gives rise to more consistency among the assignments of lexical semantic structure, reducing the degree of polysemy that is known to be too fine-grained in WordNet (Fellbaum and Grabowsky, 2002). Secondly, the method has a broad coverage since the 126 semantic types cover around 40 000 nouns. On the other hand, the results are not evaluated so we have no idea wether these 40 000 polysemic words are really instances of polysemy classes described by the 126 semantic types. Another criticism that could be addressed regarding the semantic types is that they are very general and thus not very intuitive, in that they give little information on the type of polysemy relation.

Our study is closer to the one presented in (Peters, 2006) since the author exploits both the synset hierarchy **and** glosses associated to the synsets in order to select candidates for instantiations of regular polysemy. The first step consists in an automatic selection of candidates on the basis of systematic sense distribution of nouns (pairs of hypernyms that subsume the sense combinations of the words involved). For example, in two of their senses, the nouns LAW, ARCHITECTURE, LITERATURE, POLITICS and THEOLOGY fall under the pattern `profession∼discipline`. The second stage consists in characterizing semantic relations between the related word senses by analyzing the glosses that are associated with these pairs of word senses and their hypernyms. If a verb occurs between each pair of word senses, it is taken as the semantic relation that holds between them. For example, the predicate `is mastered by` describes the semantic relation that holds between pairs of word senses that fall under the pattern `profession∼discipline`. Such relations have been extracted for around 5000 candidates. The author gives several examples of relations but doesn't say how many there are. Nor does he give an evaluation of the data that have been extracted.

3 Goal and Method

Our main goal is to describe regular polysemy relations of WordNet and to automatically detect their occurrences in the database. We then propose to enhance the lexicon by providing new lexical relations, in this case metonymy and metaphor relations.

Our method has four steps. The first one consists in extracting "auto-referent" synsets, that is, synsets whose definition includes one word that shares the form of one of the defined lexical units. The second step consists in manually describing polysemy patterns

from the observation of the 1984 synsets extracted in the first step. In the following step, we disambiguate, with the help of the patterns, the meaning of the polysemic word of the definition that is implied. The last step consists in generalizing the method to synsets that are not "auto-referent".

3.1 Extracting "auto-referent" candidates

We decided to first extract "auto-referent" synsets, that is, synsets whose definition includes one word that shares the form of one of the defined lexical units. This decision relies on the well known fact that a semantic link between two lexical units (L1 and L2) is more obvious if L2 is definied by L1. Here are a few examples of these "auto-referent" synsets extracted during the first step :

- {cerise#1, **cherry**#4} = *the red color of* ***cherries***
- {**driver**#3} = *a golfer who hits the golf ball with a* ***driver***
- {**falsify**#4} = ***falsify*** *knowingly*

Before extracting these auto-referent synsets, we automatically attribute morpho-syntactic tags to the words contained in the definitions[4] in order to make sure that L1 and L2 share the same part of speech. For instance, the two following synsets are not auto-referent since the defined lexical unit is a verb while the lexical unit used in the definition is a noun (in the case of GUN) and an adjective (in the case of SLOW)[5].

- {**gun**} = *shoot with a* ***gun***
- {**slow**#2, slow down, slow up, slack, slacken} = *become* ***slow*** *or slower*

We also eliminate (by identifying the three key words *equal*, *trade name* and *capital of*) synsets in which L1 corresponds to L2, as illustrated below :

- {kopec, **kopeck**, copeck} = *100* ***kopecks*** *equal 1 ruble in Russia*
- {sildenafil, sildenafil citrate, **Viagra**} = *virility drug name* ***Viagra***
- {Bern, Berne, **capital of Switzerland**} = *the* ***capital of Switzerland***

After this first step we obtained 1984 synsets that are likely to be occurrences of regular polysemy relations.

[4] For this task, we used the Antelope NLP framework (www.proxem.com) (Chaumartin, 2009).

[5] We have chosen to adopt this restriction to limit our first investigation of regular polysemy in WordNet. However, our method of description of regular polysemy can be applied to pairs of lexical units that share the same form but that do not belong to the same part of speach. These pairs are indeed numerous in English and are linked by regular semantic links, as illustrated above with the GUN example (manner verbs).

3.2 Describing polysemy patterns

The second step consists of manually describing polysemy patterns from the observation of the 1984 synsets extracted in the first step. The method we use to attribute a category of regular polysemy to an occurrence L1-L2 relies on two main criteria applied to WN definitions (Martin, 1972), (Fass, 1988) : the position of the inclusion of L1 in the definition of L2 and the elements from the definition that are pertinent with respect to the polysemy relation. We will first present these two criteria and explain how they interact in the attribution of a regular polysemy category to a given occurrence. We will then present some examples of polysemy patterns on which our study is based.

3.2.1 Elements of the polysemy patterns

First of all, we consider the position of L1 in the definition of L2. Indeed, L1 can be included in the first part of the definition (and then be the *genus* of the definition), as illustrated by BEHAVE below, or in the rest of the definition (and be one of the *differentiae*), as illustrated by SWEEP :

- {**behave**#3} = ***behave*** *well or properly*
- {**sweep**#6} = *clean by* ***sweeping***

In addition to the position of L1 in the definition of L2, we extract sub-strings that are recurrent in the definitions. Here are three examples of inclusion in the second part of the definition distinguished by sub-strings that introduced them :

- {**mint**#5} = *a candy that is flavored with a* ***mint*** *oil* ($\rightarrow$ "that is flavored with L1")
- {**bluefish**#2} = *fatty bluish flesh of* ***bluefish*** ($\rightarrow$ "flesh of L1")
- {**fin**#5} = *a stabilizer on a ship that resembles the* ***fin*** *of a fish* ($\rightarrow$ "that resembles L1")

Merging these two criteria (position of the inclusion and elements of the definition), one can automatically attribute a category (specialization, metaphor or metonymy) to a given occurrence. If L1 is in the first part of the definition of L2, the relation is a specialization or a metaphor. The two examples given below both show an inclusion of L1 in the first part of the definition but the first one is a specialization and the second one is a metaphor.

- {**arrange**#5} = ***arrange*** *attractively*
- {**grow**#9} = ***grow*** *emotionally*

If L1 appears in the second part of the definition, the relation is a metonymy or a metaphor, according to the element that introduces this inclusion. Among the three

examples given above, (TO) SWEEP[6], MINT, and BLUEFISH are examples of metonymy whereas FIN is an example of metaphor. The ambiguity of the category can be resolved by definitional patterns. For instance, if the inclusion is preceded by "*that resembles*", as in the definition of `fin#5`, it means it's a metaphor, not a metonymy.

3.2.2 Examples of polysemy patterns

Let us now turn to three concrete examples taken from our 60 polysemy patterns[7]. First, we present a few lines of code implementing the `colorOf` pattern :

```
patterns.Add(new Pattern("colorOf")
    .AddType("color","fruit")
    .AddType("color","gem")
    .AddType("color","metal")
    .AddMatchingRule("color of *")
```

The first line of code defines the polysemy pattern. The three following lines indicate that the pair of lexical units that are likely to be applied this pattern are of types `color` for L1 and `fruit`, `gem` or `metal` for L2. The last line says that the definition of L2 must include the string `colorOf` to be detected as an occurrence of this pattern. Another kind of pattern is the one presented below :

```
patterns.Add(new Pattern("causedBy")
    .AddMatchingRule("resulting from *")
    .AddMatchingRule("caused by *")
```

Unlike the previous pattern, this pattern does not constrain the type of L1 and L2 but only imposes that L2 contains in its definition the string `resulting from` or the string `caused by`. Indeed, the "`caused by`" relation may gather instances of too general types. For example, "`disease` is caused by `organism`" (ERGOT) and "`shape` is caused by `happening`" (DISTORTION) could only be generalized by "`something` is caused by `something`", which is too general to be relevant. As one can see, the sole exploitation of the hierarchy of synsets (that provides types for L1 and L2) would lead to missing some regular polysemy relations (detected only by patterns extracted from definitions).

The application of these polysemy patterns to the 1984 candidates allows us to extract 1427 occurrences of polysemy relations. The other 557 synsets correspond to orphan occurrences of polysemy, that is, to pairs of lexical units that are not associated with **regular** polysemy relations. For example, the pair of lexical units given below is considered as an orphan occurrence of polysemy since the metonymy pattern "`document` written on `paper`" has no instance in WordNet except PAPYRUS.

[6] Note that the verb **sweep**#6 means *to clean (in a certain way)*, not *to sweep (in a certain way)* and thus cannot be considered as an instance of specialization, as it is defined above (see section 2.1).

[7] This patterns can be found online ([8]).

- {**papyrus**#1} = *paper made from the papyrus plant by cutting it in strips and pressing it flat; used by ancient Egyptians and Greeks and Romans*

- {**papyrus**#3} = *a document written on* ***papyrus***

3.3 Disambiguating L1 in the definition of L2

At this stage, L1 is not disambiguated in the definition of L2. The three following examples are detected as occurrences of the `colorOf` pattern, but we don't know if L1 is typed `fruit`, `gem` or `metal` :

- {**emerald**#3} = *the green color of an* ***emerald***

- {tan#2, **topaz**#3} = *a light brown, the color of* ***topaz***

- {**copper**#4} = *a reddish-brown color resembling the color of polished* ***copper***

The patterns based on pairs of types are used to disambiguate L1 in the definition of L2. The method is the following : the system enumerates all the possible types for L1 (other than L2) and stops when the couple L1 L2 matches one of the pairs of types expressed by the pattern. When the situation requires it, the system explores the hierarchy of hypernyms for nouns and verbs). L1 is then disambiguated, as illustrated below :

- {emerald#3} = *the green color of an* ***emerald#1***$_{[gem]}$

- {tan#2, topaz#3} = *a light brown, the color of* ***topaz#2***$_{[gem]}$

- {copper#4} = *a reddish-brown color resembling the color of polished* ***copper#1***$_{[metal]}$

3.4 Generalizing the application of patterns

Finally, we look for new candidates by dropping the "autoreferent" definition constraints. We only rely here on the type of the synsets pair (L1, L2) provided by the patterns defined at step 2. This enables us to find 367 new instances, two of which are illustrated below. The word GOLD has five meanings in WordNet : `coins`, `color`, `metal`, `great wealth` and `something that is precious`. Two meanings are typed `color` and `metal` and can thus be associated with the rule `ColorOf`, even if the definition of *gold#2* doesn't include the word GOLD.

- {amber#1, gold#2} = *a deep yellow color* (implicit link to *gold#3*[metal])

- {coral#1} = *a variable color averaging a deep pink* (implicit link to *coral#2*[gem])

This generalization needs to be refined. Indeed, results of the generalization are good with some pattern, but bad with patterns constrained by types that are too general, like `entity`, `artefact`, `abstraction`, ...: the lack of constraints on the iclusion of L1 in the definition of L2 leads to the multiplication of candidates that do not belong to regular polysemy classes. For that reason, we require that the definition of L1 and L2 share at least some words by using a gloss overlapping similarity measure, with a TF-IDF weighting[9].

- {footstep#1} = *the sound of a **step** of someone **walking***
- {footstep#2} = *the act of taking a **step** in **walking***

This constraint improves precision, but the price to pay is a lower recall; we miss metaphor or metonymy instances that do not share any significant word. For example, our system correctly identifies the word TIGER as an instance of the metaphorical rule between an `animal`, `person` (based here on the analogy with the ferocity of the animal). But the system doesn't select it since the two definitions do not share any significant words in common, as illustrated bellow:

- {tiger#1} = *a fierce or audacious person*
- {tiger#2} = *large feline of forests in Asia having a tawny coat with black stripes*

4 Results

The results of the research reported here are of two kinds: a descriptive one with the classification of regular polysemy relations, and a methodological one, with the automatic detection of occurrences of regular polysemy relations with a rather good precision.

4.1 Lexical description : classification of regular polysemy relations

In this section, we propose a classification of regular metonymy and regular metaphor relations based on the patterns identified by our study. Note that our results also include a significant proportion of specialization instances (for example, {**pressure**#7} = *the **pressure** exerted by the atmosphere*). Nevertheless, this kind of relation is not appropriate for a hierarchical organization, since it is difficult to identify regular types for L1 and L2. Moreover, instances of specialization are already described in WN by the hypernymy relation.

For each class of regular polysemy relation presented here, we indicate in brackets the number of real instances compared to the number of detected candidates. For example,

[9] The *Term Frequency-Inverse Document Frequency* is a method of weighting usually used in text mining. This statistical measure allows us to evaluate the importance of a word in a definition. The weight increases proportionally according to the number of occurrences in the definition. The weight also varies according to the number of occurrences in all the definitions.

the regular metonymy relation "`playing card` **represents** `person` or `entity`" has six candidates, but only five of them are real instances of this relation[10] (5/6). Indeed, an `ace#2` is not a card that represents an `ace#3`, as illustrated below.

- {**ace**#2} (playing card) = *one of four playing cards in a deck having a single pip on its face.*
- **ace**#3} (person) = *someone who is dazzlingly skilled in any field*

This proportion is followed by a few examples of the given regular relation.

4.1.1 Classification of regular metonymy relations

L2 **represents** L1

→ `playing card` **represents** `person` or `entity` (5/6 ; QUEEN, KING ; TEN, NINE)

L2 **is caused by** L1

→ `fee` **is caused by** `action` (27/27 ; ADMISSION, ANCHORAGE)

→ `disease` **is caused by** `organism` (13/17 ; ERGOT, HERPES)

L2 **is produced by** L1

→ `sound` **produced by** `instrument` or `movement` or `device` (15/15 ; DRUM, WHISTLE ; SNAP ; BELL)

→ `piece of work` **is written by** `person`

→ `book` **written by** `writer` (no example in WN; it could be SHAKESPEARE)

→ `book` **written by** `prophet` (15/15 ; JOB, JEREMIAH)

→ `music` **written by** `composer` (9/9; MOZART, WAGNER)

→ L2 **is produced by** `plant` or `tree`

→ `fruit or plant material` **produced by** `tree or plant` (120/128 ; GUM, COTTON ; ORANGE, CITRUS)

→ `flower` **produced by** `plant` (50/51 ; CHRYSANTHEMUM, COTTONWEED)

→ `vegetable` **produced by** `plant` (13/13 ; BEAN, RADISH)

L2 **produces** L1

→ `business firm` **produces** `publication` (2/2 ; NEWSPAPER, MAGAZINE)

L2 **is from** L1

→ `beverage` **is from** `region` (4/4 ; CHAMPAGNE, CHABLIS)

L2 **is derived from** L1

→ L2 **is derived from** `animal`

→ `flesh` **from** `animal, fish, bird or crustacean` (303/303 ; RABBIT ; TROUT ; PHEASANT ; LOBSTER)

→ `animal skin` **from** `animal` (17/17 ; FOX, CHINCHILLA)

→ `wool` **from** `animal` (2/2 ; ALPACA, VICUNA)

→ L2 **is derived from** `leaf, plant, tree...`

→ `drink` **derived from** `leaf` (3/3 ; TEA, MATE)

→ `fiber` **derived from** `plant` (13/13 ; COTTON, FLAX)

[10] The reader should keep in mind that the candidates are selected both from the patterns and from a generalization of these patterns, generalization that relies on the semantic type of L1 and L2 (see section 3.4 above). In most cases, "wrong" instances are selected during the generalization. Nevertheless, it could be interesting to study this lack of precision more precisely.

→ wood **derived from** tree (70/70 ; BAMBOO, BALSA)
→ wine **derived from** vine (2/2 ; TOKAY, VERDICCHIO)

L2 **is about** L1
→ discipline **is about** abstraction (56/64 ; PHILOSOPHY, PHYSICS)
→ division **is responsible for** L1 (4/6 ; EDUCATION, ENERGY, TRANSPORT)
→ book **is about** person (6/6 ; JONAH, JOSHUA)

L2 **accompanies** L1
→ music **that accompanies** dance (32/32 ; POLKA, MAZURKA)

L2 **covers** L1
→ cloth covering **that covers** body part (14/15 ; ELBOW, KNEE)

L2 **is included in** L1
→ substance **included in** medicine (17/17 ; ARNICA, MENTHOL)
→ person **member of** group (37/39 ; SAMURAI, NINJA)
→ person **that is in** construction (6/6 ; BUILDING, FLOOR)
→ quantity **contained in** container (39/39 ; TEASPOON, BAG)
→ quantity of food **contained in** dish (5/5 ; PLATE, CASSEROLE)
→ river **passes by** state (6/6 ; ALABAMA, DELAWARE)
→ country **on** island (22/22 ; IRELAND, MALTA)

L2 **is typical of** L1
→ ball **for** game (7/7 ; PAINTBALL, VOLLEYBALL)
→ wine **from** region (4/4; CHABLIS, BORDEAUX)
→ color **typical of** L1 (7/7)
→ color **typical of** gem (TOPAZ, EMERALD)
→ color **typical of** metal (GOLD, COPPER)
→ color **typical of** fruit (CHERRY, CHESTNUT)
→ food **tastes** L1 (13/25)
→ food **that tastes** herb (MINT, RATAFIA)
→ part of garment **characterized by** part of the body (12/14 ; BACK, SHOULDER)
→ person **characterized by** L1
→ sportsman **characterized by** position (31/31 ; CENTER, WINGBACK)
→ singer **characterized by** voice (11/11 ; CONTRALTO, SOPRANO)

language **spoke by** person (199/223 ; KOREAN, PORTUGUESE)

4.1.2 Classification of regular metaphor relations

→ L2 **is similar to** L1
→ human communication **similar to** animal communication (3/4 ; TO BARK, TO CACKLE)
→ animal part of the body **corresponds to** human part of the body (3/3 ; LEG, THROAT)
→ person **'s behaviour resembles** animal (36/54 ; PIRANHA, POPINJAY)
→ object **'s form resembles** natural object (38/38 ; MOON, SNAKE)
→ artifact **resembles** part of the body (5/5 ; NOSE, THROAT)

4.2 Evaluation of the automatic process

As far as we know, there is no gold standard for this kind of experiment. We manually evaluated the 2351 occurrences of regular polysemy relations proposed by our system. We estimate that 2140 are correct, that is a precision of 91,03%.

We didn't find any method that allows a precise automatic evaluation of recall. However, we manually evaluated the recall for two of the regular polysemy patterns presented above: the metaphoric relation `person resembles animal` and the metonymic relation `wood derived from tree`. We manually identified 142 occurrences of the metaphors in WN (recall of 36/142=25,3%) and 79 occurrences of the metonymy (recall of 70/79=88,6%). As one can see, the recall also depends on the nature of the relation, that can be more or less regular.

5 Conclusion

In this paper we propose a method to automatically extract, with a good precision, new lexical relations in WordNet : metonymy and metaphor relations. The resource containing these new relations is available online [11]. Our results can be used in a lexical disambiguation task to infer meanings that are not described in WN. For example, in WN, the word BORDEAUX denotes both the location ({bordeaux#1} = *a port city in southwestern France*) and the wine ({bordeaux#2, bordeaux wine#1}) but BOURGOGNE denotes only the location. Our patterns could be dynamically used to create new interpretation, when the context requires it (*They convince you to drink Bourgogne*).

Our method has been tested on English WordNet but could be applied to other semantic data for the English language, for example to FrameNet data (Fillmore et al., 2003). Indeed, FrameNet also provides an hierarchy of concepts that can be expressed by lexical units (the frames) **and** definitions that analyse the meaning of these lexical units. It could be interesting to compare the result of our method applied on WordNet with those of the same method applied to FrameNet, to see which regular polysemy patterns they share and to see which instances we find for each pattern.

Our method can also be applied to WordNets that are devoted to other languages (once the patterns description is adapted to the language) to help the systematic encoding of definitions and to compare regular polysemy relations that are shared by different languages. This however involves that the ressources propose definitions for each lexical units.

References

Apresjan, J. (1974). Regular Polysemy. *Linguistics*, 142:5–32.

Apresjan, J. (1992). *Lexical semantics*. Karoma Publisher, Ann Arbor.

Buitelaar, P. (1998). Corelex : An ontology of systematic polysemous classes. In *Proceedings of FOIS98, International Conference on Formal Ontology in Information Systems*, Amsterdam.

Chaumartin, F.-R. (2009). Antelope : une plateforme industrielle de traitement linguistique. *Traitement Automatique des Langues*, 49.

Copestake, A. and Briscoe, T. (1995). Semi-productive Polysemy and Sense Extension. *Journal of Semantics*, 1:15–67.

Fass, D. (1988). Metonymy and Metaphor : What's the difference? In *Proceedings of Coling-88*, pages 177–181.

[11] http://www.chaumartin.fr/download/wnpolysemy.zip

Fellbaum, C. (1998). *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, Mass.

Fellbaum, C. and Grabowsky, J. (2002). Polysemy and the (mental) lexicon. In Lenci, A. and Tomaso, V. D., editors, *Exploring the lexicon. Theory and Computation*. Edizioni dell'Orso, Alessandria.

Fillmore, C., Johnson, C., and Petruck, M. (2003). Background to Framenet. *International Journal of Lexicography*, 16:235–250.

Johansen, J. and Larsen, S. (2002). *Signs in Use. An introduction to semiotics*. Routledge, London/New-York.

Martin, R. (1972). Esquisse d'une analyse formelle de la polysémie. *Travaux de linguistique et de littérature*, 10:125–136.

Ostler, N. and Atkins, B. (1991). Predictable meaning shift : Some linguistic properties of lexical implication rules. In Pustejovsky, J. and Bergler, S., editors, *Lexical Semantics and Knowledge Representation : First SIGLEX Workshop Proceedings*. Springer-Verlag, Berlin.

Peters, W. (2006). In Search for More Knowledge : Regular Polysemy and Knowledge Acquisition. In *Proceedings of GWC2006*.

Pustejovsky, J. (1995). *The Generative Lexicon*. MIT Press, Cambridge.

Ravin, Y. and Leacock, C. (2000). Polysemy: An overview. In Ravin, Y. and Leacock, C., editors, *Polysemy, Theoretical and Computational Approaches*. Oxford University Press, Oxford.

Nofiza Vokhidova

Überlegungen zur Erweiterung lexikalisch-semantischer Ressourcen durch die Graduonymie

Diese Arbeit untersucht das Phänomen der Graduierung im Bereich der lexikalischen Semantik[1]. Im lexikalischen System einer Sprache existieren Wörter, die durch verschiedene Grade eines Merkmals kontrastieren und eine besondere lexikalisch-semantische Gruppe konstituieren. Die Relation zwischen den Elementen derartiger Gruppen wird als Graduonymie bezeichnet. Diese Studie stellt Überlegungen zur Erweiterung lexikalisch-semantischer Ressourcen durch die Graduonymie dar. Insbesondere wird untersucht, wie graduonymisch aufeinander bezogene Wörter in lexikalischen Online-Ressourcen repräsentiert sind und welchen Stellenwert diese Relation in solchen Systemen aufweisen könnte. Durch einen Wörterbuchabgleich wird die Repräsentation der graduonymischen Paradigmatik von KIND im Online-Wörterbuch *elexiko* überprüft und mittels Korpusanalysen näher bestimmt. Es handelt sich dabei um eine korpusbasierte Untersuchung mit sowohl quantitativem als auch qualitativem Ansatz.

1 Einleitung

Die Erforschung paradigmatischer Beziehungen zwischen lexikalischen Zeichen ist von den Ursprüngen bis zu jüngsten computergestützten lexikologisch-lexikographischen Arbeiten[2] nach wie vor höchst relevant. Die bisher auf klassische Sinnrelationen wie Synonymie und Antonymie beschränkten Relationsarten des lexikalischen Systems sind in den computerverfügbaren und webbasierten lexikalisch-semantischen Ressourcen durch weitere Sinnrelationen angereichert worden, z.B. Hyperonymie/Hyponymie, Parteronymie/Partonymie (auch: Holonymie/Meronymie), die als vertikale Struktur paradigmatischer Muster (vgl. Storjohann, 2005b nach Cruse, 2004, 1986; Lutzeier, 1981) klassifiziert werden und neben prominenten Sinnrelationen wie der Synonymie und Antonymie, Inkompatibilität, Komplementarität, Konversonymie, Reversität, Plesionymie, Heteronymie, Troponymie, Pertinenzrelationen, die horizontale Struktur paradigmatischer Muster aufweisen. Diese Arbeit hat zum Ziel, Überlegungen zur Anreicherung horizontaler Strukturen paradigmatischer Relationen durch einen weiteren Relationstyp darzustellen. Es handelt sich bei diesem Relationstyp um die Beziehung zwischen den

[1] In dieser Arbeit werden Teilaspekte eines Dissertationsprojekts vorgestellt.

[2] Zu den lexikalisch-semantischen Datenbanken gehören beispielsweise WordNet, EuroWordNet, GermaNet, UniNet, FrameNet; zu den Wortschatz-Informationssystemen, die sich mit der Paradigmatik der Ausdrücke beschäftigen, gehören das Wortschatz-Portal, *elexiko*, *ordnet.dk*, ANW, ELDIT etc. (Für diese Arbeit relevante Online-Ressourcen werden am Ende des Artikels aufgelistet).

Wörtern, die aufgrund gradueller Ausprägung eines spezifischen Merkmals in ihrer semantischen Struktur in Kontrast stehen und sich dadurch den lexikalisch-semantischen Gruppen zuordnen lassen. Dieses Phänomen wird als *Graduonymie*[3] bezeichnet.

2 Graduonymie als Sinnrelation – ein Überblick

Wie andere Sinnrelationen manifestiert sich Graduonymie zwischen lexikalischen Einheiten im Wortschatz und strukturiert lexikalische Paradigmen, deren Elemente auf Grund gradueller Unterschiede in ihrer Bedeutung kontrastieren. Diesem semantischen Unterschied liegt eine stufenweise Steigerung oder Verringerung eines spezifischen Merkmals in der semantischen Struktur der Wörter – *Einzelbedeutungen von Wörtern*, genauer *Lesarten* oder *Sememe* – zugrunde. Aufgrund gradueller Ausprägungen von semantischen Merkmalen dieser Wörter werden graduelle Ketten systematisiert. Mögliche Bezeichnungen für eine graduelle Kette sind: *linguistische Skala*, *graduonymisches Paradigma, graduonymische lexikalisch-semantische Gruppe, graduonymische Reihenfolge*. Die Glieder eines graduonymischen Paradigmas werden *Graduonyme* genannt. Ab einer Reihe von mindestens drei Gliedern ist Graduonymie präzise zu erkennen: Pforte >> Tür >> Tor. Eine graduonymische Kette kann aber auch mehr als drei Glieder umfassen: Neugeborenes >> Säugling >> Kind >> Jugendlicher >> Erwachsener. Die Glieder eines graduonymischen Paradigmas formieren sich oft um ein dominantes Wort (Zentrum, Kern) herum, welches die Bedeutung anderer Wörter subsumiert und den neutralen Grad des steigerungsrelevanten Merkmals aufweist. Generell wird in der Literatur das Kernwort des lexikalischen Paradigmas als Oberbegriff bezeichnet und es werden je nach der Art des Paradigmas der Sinnrelationen spezielle Termini verwendet[4]. Zum Beispiel wird ein Oberbegriff (übergeordneter Begriff) für untergeordnete Begriffe einer Hyponymiebeziehung als *Hyperonymie* oder *Superordination* (Lyons, 1980, S. 301) bezeichnet. Für die Gesamtheit einer hierarchischen Partonymie- oder Meronymie-Relation, in der einige Teile in ein Ganzes eingeschlossen werden, wird der Terminus *Parteronym* (Lutzeier, 1995, S. 76 nach Wiegand, 1998, S. 911) oder *Holonym* (Cruse, 1986, S. 162) verwendet. Ein Wort mit stilistisch neutralen Nebenbedeutungen einer Synonymgruppe wird *Grundsynonym, Leitsynonym* oder *Dominante* (Agricola (1982, S. 19), Filipec (1966, S. 270), Schippan (2002, S. 212)) genannt. Analog zu diesem Prinzip wird die Auffassung vertreten, für das vorherrschende Wort einer graduonymischen Reihenfolge einen speziellen Fachausdruck einzuführen. Insofern wird in dieser Arbeit für den Begriff, der in einer graduonymischen Gruppe durch einen neutralen Grad eines spezifischen Merkmals gekennzeichnet wird und als Oberbegriff fungiert, der Terminus *Hypergraduonym* verwendet. Die umgebenden Elemente einer graduonymischen Reihenfolge werden je nach dem Steigerungsgrad des Merkmals dem Hypergra-

[3] Ein erster Überblick zur strukturalistischen Erforschung des Phänomens der Graduonymie findet sich in: Ne'matov et al. (1989); Aripzhonova (1994); Ne'matov et al. (1995); Vokhidova (2007).

[4] Dieses Ordnungsprinzip ist zudem für Wortfelder zutreffend. In der Wortfeldtheorie wird ein Lexem, *„dessen Inhalt mit dem eines ganzen Wortfeldes identisch ist“*, als *Archilexem* bezeichnet (Coseriu, 1970, S. 112).

duonym zugeordnet. Während die Wörter auf der rechten Seite des Hypergraduonyms die Steigerung eines Merkmals signalisieren, zeigen die Wörter auf der linken Seite die Verringerung dieses Merkmals auf (Abb. 1).

Abbildung 1: Stellenwert von WIND als Hypergraduonym in der Gruppe

Das allgemeine Ordnungsprinzip des lexikalischen Paradigmas ist auch für lexikalische Graduonymie zutreffend:

- Ein lexikalisches Paradigma wird aufgrund einer für alle Glieder gemeinsamen Bedeutung konstituiert;
- Es bestehen spezielle Unterschiede zwischen den Gliedern eines Paradigmas neben ihren identischen Bedeutungskomponenten;
- Im Paradigma dominiert ein denotativ und/oder konnotativ neutraler oder übergeordneter Begriff.

In der Abbildung 2 wird dies anhand der oben angeführten graduonymischen Reihenfolge veranschaulicht. Die Wörter Brise, Wind, Sturm und Orkan haben die gemeinsame Bedeutung *'Luftbewegung'*. Die Verschiedenheit der Wörter besteht in dem differenten Grad der *'Windstärke'*. Dieses Merkmal ist für diese Gruppe ein steigerungsrelevantes (graduierbares) Merkmal und es steigert sich von der *Brise* bis zum *Orkan*. *Wind*, wie es die Abbildungen (Abb. 1 und 2) illustrieren, fungiert in diesem Paradigma als Hypergraduonym.

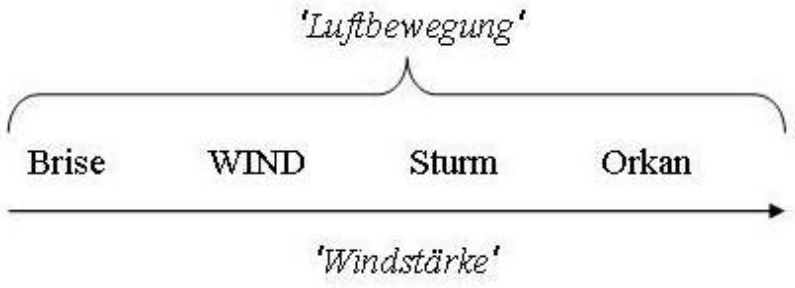

Abbildung 2: Verhältnis der Wörter zueinander in der Paradigmatik von WIND

Das Hypergraduonym ist gleichzeitig ein gradueller (inkompatibler) Partner zu anderen Elementen einer graduonymischen Reihe. Es fungiert je nach Kontext sowohl als

Hypergraduonym, als auch als *Graduonym*. Die folgenden Kospusbeispiele zeigen diese spezifische Eigenschaft eines Hypergraduonyms am Beispiel von *Wind*.

(1) **Wind** kann selbst dann Schrecken verbreiten, wenn er als laues Lüftchen säuselt. Als Ende Februar in Großbritannien die Maul- und Klauenseuche ausbrach, konnte schon eine leichte **Brise** Landwirte ängstigen, deren Höfe nahe bei verseuchten Tierherden lagen. (Mannheimer Morgen, 04.10.2001, Ressort: Welt und Wissen; Der Wind bringt Bewegung in die Natur - aber das ist nicht immer ein Grund zur Freude)

(2) Sie wurden von Generationen von Golfern und grasenden Kühen geformt. Hier bläst ständig eine frische **Brise**, oft ein scharfer **Wind** aus immer wieder wechselnden Richtungen, der die Bälle in weitem Bogen vom Kurs abtreibt. Und hier bestimmt der Caddie, wie das Spiel gespielt wird. (Die Zeit, 19.07.1985, S. 42; Ein Birdie an Loch sechs)

Für eine graduonymische Reihe ist aber ein Oberbegriff in Form eines Lexems (lexikalisiertes Wort) nicht obligatorisch (vgl. Lyons, 1980; Ne'matov et al., 1995; Blank, 2001). D.h., eine Graduonymiereihung kann auch ohne Hypergraduonym existieren. Im Falle solcher lexikalischen Lücken werden die Paradigmen durch andere sprachliche Mittel gefüllt (siehe z.B. (a) und (c) im Kapitel 6).

Anhand verschiedener Merkmale kann für Substantive, Adjektive, Verben, Adverbien, Pronomen eine Vielzahl graduonymischer Reihenfolgen nachgewiesen werden. Da aufgrund der graduellen Opposition keine lexikalischen Paradigmen im Bereich der Sinnrelationen speziell differenziert worden sind, werden solche Einheiten, die als Graduonyme interpretierbar sind, in der Lexik bei den synonymischen oder hyponymischen lexikalisch-semantischen Gruppen aufgeführt oder sie sind als einzelne Wörter definiert, ohne explizite Berücksichtigung der lexikalisch-semantischen Relationen. Der nächste Teil der Arbeit ist dem Stand der Forschung zur Graduonymie in der Literatur gewidmet.

3 Zum Stand der Forschungen zur Graduonymie

Das Phänomen der Graduierung stellt ein aktuelles Problem in der linguistischen Forschung dar, das sich auf verschiedenen Betrachtungsebenen manifestiert. Zu den gängigen Graduierungsphänomenen in Grammatik und Lexikon gehört die graduelle Abstufung eines Merkmals auf der lexikalischen Ebene. Hierbei wird die Bedeutung von Merkmalen und Eigenschaften durch verschiedenartige lexikalische Mittel graduiert dargestellt. Intensivierung des Merkmals durch Modifikatoren, namentlich durch Gradpartikeln (Steigerungspartikeln, Intensitätspartikeln) ist in der Literatur eingehend untersucht. Es gibt im lexikalischen System der Sprache eine andere Art der Graduierung, die mit der semantischen Struktur der lexikalischen Einheiten verbunden ist. Der Grad der Stärke/Schwäche eines Merkmals manifestiert sich in einem anderen Wort, mit dem

es inhaltlich assoziiert ist. Es entwickelt sich daraus eine Kette von Wörtern, die unterschiedliche Grade eines Merkmals signalisieren. Überlegungen zur Steigerung oder Verringerung semantischer Merkmale in graduellen Ketten sind in der Sprachwissenschaft nicht unbekannt. Im Folgenden werden die wichtigsten Arbeiten im Bereich erwähnt, welche die Phänomene untersuchen, die als Graduonymie interpretiert werden können.

Die Graduierung auf der lexikalischen Ebene der Sprache zwischen den Wörtern mit ähnlicher gemeinsamer denotativer Bedeutung und die daraus hervorgehenden lexikalischen Skalen gehen auf Untersuchungen Coserius zurück, wodurch die strukturalistische Hypothese über die Existenz der zunächst im phonologischen System entwickelten Oppositionsarten auf das lexikalische System der Sprache übertragen wurde (Coseriu, 1978, S. 64). Ein typisches Beispiel dazu ist die adjektivische Temperaturbezeichnung *eisig – kalt – kühl – lau – warm – heiß*. Die Relation zwischen lexikalischen Einheiten bezieht sich auf die graduelle Opposition. Diese Auffassung wurde in der Wortfeldtheorie für einzelne Wortfelder[5] untersucht, die auf graduellen Bedeutungsbeziehungen aufgebaut sind.

Lyons unterteilt die Mengen von inkompatiblen Ausdrücken in Serien und Zyklen. Bei seriell geordneten Mengen unterscheidet Lyons weiterhin in Bezug auf die Graduierbarkeit der konstituierenden Lexeme *Skalen* und *Grade* (vgl. Lyons, 1980, S. 299). Skalen wie {‚heiß', ‚warm', ‚kühl', ‚kalt'} und {‚ausgezeichnet', ‚gut', ‚mittelmäßig', ‚dürftig', ‚schlecht', ‚misarebel'} werden hier als Graduonyme betrachtet.

Unter den Untersuchungen im Bereich der lexikalischen Semantik sind die Ansätze von D.A. Cruse von besonderer Relevanz (1980; 1986; 2002; 2004). Cruse untersucht lexikalische Ausdrücke, die als Graduonyme bezeichnet werden können, unter verschiedenen Aspekten, die im Folgenden genannt werden.

Als speziellen Typ der Synonymie bezeichnet Cruse die Relation zwischen semantisch dicht beieinander liegenden Ausdrücken mit dem Terminus **Plesionymie** (Cruse, 1986 nach Storjohann, 2006). Im Vergleich zu den absoluten und propositionalen Synonymen, die eine bedeutungsgleiche und bedeutungsähnliche Verwendung aufweisen, werden Plesionyme (engl. *near-synonymy* oder *parasynonymy*)[6] (z.B. *diesig - dunstig - nebelig*) durch ihren kontrastiven Gebrauch gekennzeichnet. Dies bestätigt zudem Storjohann in ihrer Studie zur kontextuellen Variabilität synonymer Relation anhand ihrer Korpusbeobachtungen:

> Während bei einigen bedeutungsgleichen Ausdrücken die synonymen Kontexte wesentlich häufiger belegt sind, gibt es auch Wortpaare, die ein ausgewogenes Verhältnis zwischen beiden Möglichkeiten aufweisen, und es gibt Synonyme, bei denen der kontrastive Gebrauch stärker im Vordergrund steht. In Fällen, in denen der kontrastive Gebrauch stärker belegt ist als der synonymische, handelt es sich vor allem um Plesionyme. (Storjohann, 2006, S. 11)

[5] Zum Beispiel: Wortfelder „Gewässer", „Hörbare Schwingung", „Verben der Fortbewegung" etc.
[6] Cruse (2002, S. 490).

Nach Cruse und Storjohann sind die Einträge der Synonymwörterbücher zu einem großen Teil Plesionyme (vgl. Storjohann, 2006, S. 11).

Absolute Synonymie liegt dann vor, wenn zwei Ausdrücke bei gleichbleibender Bedeutung in allen Kontexten austauschbar sind. Bei der propositionalen (kognitiven) Synonymie handelt es sich um lexikalische Einheiten mit gleicher Denotation, aber mit unterschiedlicher expressiver Bedeutung (Konnotation), z.B. *Geige* und *Violine*. Plesionyme drücken hingegen einen Bedeutungsgegensatz in der denotativen Seite der Lexeme aus, z.B. *klein - winzig*; dieser Bedeutungsgegensatz führt häufig zur Intensivierung der Merkmale (vgl. Storjohann, 2006, S. 12). Die Besonderheiten zwischen drei unterschiedlichen Arten der Synonymie können unter der Annahme einer Peripherie und eines Zentrums in einer synonymischen Gruppe konkretisiert und fundiert werden[7].

- Wenn zwischen zwei Ausdrücken absolute Synonymie besteht, haben beide Ausdrücke ohne Funktion eines Zentrums und einer Peripherie die gleiche Position in der Gruppe: {L1 = L2} (Lexem1 und Lexem2 sind inhaltlich identisch).

- Bei den propositionalen Synonymen fungiert ein Lexem mit stilistisch neutralen Nebenbedeutungen, umfangreicher Distribution und größer Häufigkeit als Zentrum (Dominante) der Gruppe. Als peripheres Synonym wird ein Wort mit unterschiedlichen konnotativen, regionalen, historischen, sozialen Bewertungen bezeichnet (vgl. Schippan, 2002, S. 212). Beide Wörter weisen gemeinsame relevante Merkmale auf; der Unterschied zwischen ihnen liegt in den zusätzlichen Nebenbedeutungen des peripheren Synonyms. Die Relation zwischen Peripherie und Zentrum kann anhand folgender Konstruktion veranschaulicht werden: {Peripherie = Zentrum (Dominante) + Konnotation}. Dementsprechend: {Antlitz = Gesicht + gehoben}.

- Die Verschiedenheit zwischen Plesionymen besteht darin, dass sie durch ihre unterscheidenden Merkmale in der Denotation kontrastiert werden. Der Bedeutungsgegensatz ist mit dem Grad eines semantischen Merkmals verbunden. Diese Relation unterscheidet sich eindeutig von absoluten und propositionalen Synonymen. Auf Intensivierung bezogene Plesionyme werden in dieser Arbeit Graduonyme genannt. Die Konstruktion von Graduonymen ist: {Peripherie = Zentrum (Hypergraduonym) + Grad}. Dementsprechend: {Sturm = Wind + Stärke}.

Daraus ergibt sich, dass man bei der Synonymie auf drei verschiedenartige Phänomene stößt.

Cruse unterscheidet bei lexikalischen Konfigurationen Taxonomien, Teil-Ganzes-Hierarchien und nicht verzweigende Hierarchien als bestimmte Typen einer Hierarchie (1986, S. 122 ff.). Als Subtyp einer nicht verzweigenden Hierarchie werden zwei grundlegende Skalen (192 ff.) unterschieden:

[7] *„Synonymgruppen haben den Charakter von Teilsystemen und sind damit im Sinne der Prager Schule durch Zentrum und Peripherie gekennzeichnet. [...]; die Annahme von Peripherie und Zentrum hat auch im Bereich der Synonymik volle Gültigkeit“* Schippan (2002, S. 211-212).

- Lexikalische Einheiten, die auf einer diskontinuierlichen Skala angeordnet werden, werden *rank-terms* (*major, colonel, ...* ; *first, second, ...* ;) genannt. Für solche Begriffe bieten sich keine Graduierung an.

- Begriffe, die eine kontinuierliche Skala bilden, können graduierbar oder nicht graduierbar sein; die nicht graduierbaren Begriffe werden *degree-terms*, die graduierbaren werden *grad-terms* genannt.

Sowohl *degree-terms* (*baby, child, adolescent, adult*), die eine temporale Aufeinanderfolge darstellen, oder Ausdrücke, die durch das Merkmal *Größe* abgestuft werden (*mound, hillock, hill, mountain*), als auch *grad-terms* (*freezing, (cold), cool, warm, (hot), scorching* oder *minuscule, tiny, (small), (big), huge, gigantic*) werden in dieser Studie als Ergebnisse der Konstruktion {Peripherie = Zentrum (Hypergraduonym) + Grad} als *Graduonyme* interpretiert.

Trotz der Arbeiten in der linguistischen Forschung haben die aufeinander graduell bezogenen lexikalischen Einheiten in der lexikalisch-semantischen Forschung bisher nicht den Status einer anerkannten Sinnrelation erlangt. Diese Arbeit untersucht das Phänomen unter lexikalisch-semantischem Aspekt und zielt darauf, die Relation der Graduonymie als eigenständigen Typ der Inhaltsrelationen zu etablieren.

4 Graduonymie in der Lexikographie

Sowohl in der Linguistik als auch in der Lexikographie gibt es für untereinander graduell verbundene Wörter keinen eigenständigen Relationstyp. In den Printwörterbüchern, in denen vor allem synonymische und antonymische Beziehungen der Wörter behandelt werden, werden die Wörter mit gradueller Bedeutungsbeziehung innerhalb einer Synonymgruppe (1, 2) und als Antonyme, die Zwischenbereiche signalisieren (3, 4, 5), repräsentiert:

(1) *Pforte, Tür, Tor, Portal* (Duden, 2004)

(2) *laufen, rennen, rasen, sausen* (Görner and Kempcke, 2005)

(3) *hell : [dämmrig] : dunkel* (Agricola and Agricola, 1992)

(4) *beginnen : [andauern] : enden* (Agricola and Agricola, 1992)

(5) *glühend : kochend : heiß: warm : lau : kühl : kalt : eiskalt* (Agricola and Agricola, 1992)

Auch in computer- und web-verfügbaren lexikalisch-semantischen Ressourcen, in denen die Paradigmatik von Wörtern ausführlich behandelt wird, sind die graduell aufeinander bezogenen Wörter innerhalb einer synonymischen (1, 2) oder einer hyper/hyponymischen Relation (3) angegeben:

(1) *lachen, kichern, totlachen* (Wortschatz-Portal)

(2) *weinen, wimmern, heulen* (Wortschatz-Portal)

(3) KIND: *Neugeborenes, Säugling, Kleinkind, Schulkind* (GermaNet)

Besonders interessant ist die Beschreibung der Paradigmatik im deutsch-italienischen Lernwörterbuch ELDIT. Die Graduonyme werden in diesem Informationssystem als *Troponyme*[8] behandelt und werden innerhalb einer hypero-/hyponymischen Relation repräsentiert, welche die spezielle Art eines Hyperonyms beschreiben. Zudem werden sie nach bestimmten semantischen Merkmalen systematisiert und farblich differenziert dargestellt, sowie anhand der Kommentare und Belege dokumentiert. Neben den Gruppen VIEL_SPRECHEN (*plappern, quasseln, schwatzen*, etc.), UNDEUTLICH_SPRECHEN (*nuscheln, lispeln, stottern*, etc.), UNNATÜRLICH_SPRECHEN (*säuseln, flöten*) zum Hyperonym SPRECHEN, wird auch eine Gruppe angelegt, die sich auf die Lautstärke bezieht:

(1) LAUT_SPRECHEN: *schreien, brüllen, rufen*

(2) LEISE_SPRECHEN: *flüstern, wispern, murmeln, zischen, brummen* (ELDIT-Wörterbuch)

Erwähnenswert ist die Repräsentation von Graduonymen im Online-Wörterbuch *elexiko*. Die miteinender graduell verbundenen Wörter werden in *elexiko* folgenderweise dargestellt:

- Aufgrund der Variabilität von Sinnrelationen treten lexikalische Ausdrücke in unterschiedliche Beziehungen zueinander (vgl. Lutzeier, 1995 und Storjohann, 2006). Aus diesem Grund werden sie innerhalb einer synonymischen oder partonymischen Gruppe behandelt:

  (1) als synonymische Partner: *Wind* und *Sturm*[9]

  (2) als partonymische Partner: *Wind* zu *Sturm*

- Als Beziehung der Inkompatibilität:

  (3) *Kind*: *Neugeborenes, Säugling, Jugendlicher, Erwachsener*

- Als Sonstige Beziehung(en), die dem traditionellen Klassifikationssystem der Sinnrelationen nicht zugeordnet werden können:

  (4) *Wind, Sturm, Orkan*

  (5) *mögen/gern haben, lieben*

[8] *Troponymie* ist semantische Relation der Spezifizierung, die nur für Verben definiert wird (vgl. Heusinger, 2004, S. 192).

[9] Zur näheren Erläuterung einzelner Beispiele siehe die Beschreibung des *Wind*-Artikels auf der *elexiko*-Seite.

Die Beziehung der semantischen Steigerung bzw. Graduierung wird neben kausalen und konditionalen Beziehungen innerhalb dieser Gruppe eingeordnet und anhand der Kommentare erläutert.

Aus der oben durchgeführten Analyse wird deutlich, dass die Graduonymie in der Lexikographie in verschiedenen anderen Bedeutungsbeziehungen dargestellt ist.

5 Korpusverfahren bei der Analyse der einzelnen graduonymischen Reihenfolgen

Die empirische Grundlage dieser Arbeit bildet *das Deutsche Referenzkorpus* (kurz: DeReKo) des IDS in Mannheim, das über das Programm COSMAS II[10] (Corpus Search, Management and Analysis System) zugänglich ist. Es umfasst über 3,4 Milliarden Wörter (Stand 28.07.2008) aus geschriebenen deutschsprachigen Texten der Belletristik, Sach- und Fachsprache, eine große Zahl von Zeitungstexten und weiterer Textsorten, die kontinuierlich ergänzt werden. Daten, die in dieser Untersuchung als empirische Grundlage in Korpora überprüft werden sollten, bilden vor allem die von mir erstellten Wortlisten (Reihenfolgen) zur Graduonymie. Den Ausgangspunkt für die gewonnenen graduonymischen Reihenfolgen bilden Auswertungen von Printwörterbüchern des Deutschen[11] und Angaben deutscher Informanten. Insgesamt wurden mehr als 200 Graduierungsreihen zusammengestellt. Anhand von 71 sortierten Graduierungsreihen aus dem Gesamtmaterial wurde eine Online-Befragung durchgeführt (siehe darüber Vokhidova, 2009). Nur entsprechend ausgewählte graduonymische Reihen, die in der Befragung von Muttersprachlern beurteilt wurden, werden mittels Korpusverfahren untersucht. Im Rahmen dieses Beitrags wird die graduonymische Paradigmatik von KIND analysiert. Bevor die Verifikation der graduonymischen Reihenfolgen in Korpora erfolgt, werden sie zunächst mit den Angaben eines Online-Informationssystems abgeglichen. Es handelt sich bei diesem Informationssystem um ein Online-Wörterbuch der deutschen Gegenwartssprache *elexiko* des Portals für wissenschaftliche, korpusbasierte Lexikografie OWID[12] (*Online-Wortschatz-Informationssystem Deutsch*) des IDS in Mannheim. Durch Abgleich der graduonymischen Reihenfolgen mit den *elexiko*-Angaben wird ermittelt, welche Arten der Sinnrelationen für miteinander graduonymisch verbundene Wörter in *elexiko* repräsentiert sind. Im Anschluss daran werden sowohl die ausgewählten Graduierungsreihen als auch die *elexiko*-Daten in authentischen Textsammlungen daraufhin überprüft, inwieweit sie miteinander übereinstimmen und worin die Unterschiede bestehen. Die Erfassung graduonymischer lexikalisch-semantischer Gruppen in computer- und web-verfügbaren lexikographischen Ressourcen zeigt auf, dass das *elexiko*-Wörterbuch mit seiner ausführlichen Behandlung paradigmatischer Bedeutungsbeziehungen auf graduelle Relationen zwischen Wörtern in Belegen und Kommentaren lediglich verwiesen hat. Für die Graduonymie wurde in *elexiko* kein eigenständiger

[10] Es wurde für diese Untersuchung betriebssystemunabhängige WWW-Applikation COSMAS IIweb benutzt. COSMAS II ist unter http://www.ids-mannheim.de/cosmas2/ aufrufbar.

[11] Zum Beispiel: Wahrig (1986); Duden (1986); Görner and Kempcke (2005); PC-Bibliothek (2001); Rachmanov (1983).

[12] http://www.owid.de/

Relationstyp etabliert, sondern innerhalb verwandter Sinnrelationen oder als 'Sonstige Beziehungen' beschrieben[13]. Für die Erarbeitung der Wortartikel wird in *elexiko* das spezielle *elexiko*-Korpus genutzt, das nach formalen und inhaltlichen Kriterien aus dem Deutschen Referenzkorpus des IDS Mannheim zusammengestellt wurde. Das *elexiko*-Korpus verfügt über 1,3 Milliarden Wörter aus Zeitungs- und Zeitschriftentexten, die regelmäßig erweitert werden. Dieses Korpus ist nicht frei zugänglich. Für die Analyse und Extrahierung der graduonymischen Reihenfolge wurde, wie oben erwähnt, *das Deutsche Referenzkorpus* mit allen öffentlichen Korpora des Archivs der geschriebenen Sprache (das Hauptarchiv) genutzt. Es besteht also ein wesentlicher Unterschied zwischen der empirischen Grundlage des *elexiko*-Wörterbuchs und der für diese Arbeit genutzten Korpora in Hinblick auf den Umfang. Insofern können die hier aufgeführten Frequenzdaten von den Angaben im *elexiko*-Korpus variieren.

5.1 Annäherung an die Methodik

Die Hypothese, dass es in der Sprache Wörter gibt, die sich durch ein spezifisches Merkmal in ihrer semantischen Struktur graduell aufeinander beziehen, wurde teilweise durch die Erstellung der graduellen Reihenfolgen mittels in den Wörterbüchern kodifizierten Wissens, aber auch durch die Angaben deutscher Muttersprachler in einer Online-Umfrage bestätigt. Inwieweit die gewonnenen Daten stabil und exhaustiv sind, ist noch klärungsbedürftig. In dieser Hinsicht stellt sich zu Recht die Frage, was Korpusverfahren für die Verifikation der Graduonyme leisten können. Die hier angewandten Verfahren sollten als Ergänzungsmethode, als Zusatzmaterial für die bisher durch die anderen Methoden erhobenen Daten dienen. In diesem Zusammenhang sind diese Verfahren analog zum angloamerikanischen korpuslinguistischen Ansatz als „*corpus-based*“ (vgl. Storjohann, 2005a, S. 252 ff.) oder als „*korpusgestützter, qualitativer Ansatz*“ (vgl. Lemnitzer and Zinsmeister, 2006, S. 32 ff.) zu bezeichnen. Bei diesem Verfahren handelt es sich in Bezug auf Graduonymie nicht um die Entdeckung neuer graduonymischer Reihungen, sondern um die Verifizierung, Quantifizierung der existierenden Daten und ihre Illustirierung durch Korpusbelege (vgl. Storjohann, 2005a, S. 252 ff.). Zu erwarten sind von den Ergebnissen der korpusbasierten Untersuchung die empirische Fundierung der bereits vorhandenen Datensammlung einerseits, die Ergänzung dieser Reihen durch Elemente, die der bisherigen Datenerhebung entgangen sind oder Heranziehen der „echten“ Graduonymen in die Reihe, andererseits. Hier angewandten Korpusmethoden sollen helfen, die Stabilität der graduonymischen Reihung zu überprüfen.

Es wurden in dieser Arbeit folgende Korpusanalyseschritte vorgenommen:

Erstens, mithilfe einer gezielten Suchanfrage anhand der treffereinschließenden Abstandsoperatoren, nämlich Wortabstands- und Satzabstandsoperatoren wurden die semantischen Relationen zwischen ausgewählten Suchwörtern in entsprechenden Abständen und Reihenfolgen ermittelt.

[13] Siehe: Kapitel 4.

Zweitens, mittels einer automatischen Kookkurenzanalyse wurde Kookkurenzprofil der Suchbegriffe überprüft. Durch diese Extraktionsverfahren können die Kookkurenzpartner des Suchwortes als paradigmatische Partner, unter anderem als Graduonyme, ermittelt werden. Darüber hinaus gibt die Kookkurenzanalyse einen Überblick über die kontextuellen Gebrauchbesonderheiten zum Schlüsselwort, in welcher stilistischen Umgebung es vorkommt. Diese ist vor allem für das Heranziehen stilistisch neutraler Ausdrücke in eine graduonymische Reihenfolge wichtig.

Drittens, eine KWIC[14]-Ansicht gibt eine erste Vorstellung zu Graduonymen, welches Verhältnis zwischen den Wörtern in einem kurzen Abstand innerhalb eines Kontextes vorliegt. Von besonderem Interesse ist dabei zu beobachten, anhand welcher syntagmatischen Indikatoren die Wörter in Kontrast zueinander stehen.

Viertens, manche feinen semantischen Unterschiede zwischen den Wörtern, die durch eine KWIC-Ansicht nicht eindeutig zum Vorschein kommen, lassen sich in einem Gesamt-Volltext erkennen. Graduonyme befinden sich zudem nicht immer in kontextuell unmittelbarer Nähe zueinander. Sie tauchen häufig in einer Umgebung auf, die bis drei Sätze umfasst. Auch für solche Fälle ist die Untersuchung der Volltexte mit größerem Kontext wichtig.

Fünftens, um zu ermitteln, wie oft die Ausdrücke als Graduonyme vorkommen und welche semantische Relation zwischen Ausdrücken im Vordergrund steht, sind die Angaben von Frequenzen für die Verifikation der Graduonyme wichtig. Es muss hierbei erwähnt werden, dass manche für eine graduonymische Gruppe relevanten Wörter in den Textsammlungen sehr selten auftreten.

6 Die Analyse der Paradigmatik von KIND

KIND hat in *elexiko* drei Lesarten. Von Interesse ist für uns die erste Lesart 'sehr junger Mensch'. Die Wörter einer ursprünglichen graduonymischen Reihenfolge *Erwachsener, Jugendlicher, Neugeborenes* und *Säugling* sind in *elexiko* als inkompatible Partner zu *Kind* aufgeführt und die kontrastive (inkompatible) Beziehung der Ausdrücke ist mittels Korpusbelegen dokumentiert. Wenn wir auf die semantische Struktur und die Beziehung der Wörter zueinander achten, so wird deutlich, dass zwischen den Wörtern gewisse graduelle Unterschiede bestehen, die die verschiedenen Altersmerkmale signalisieren. Dass es sich um graduelle Heterogenität zwischen den Wörtern handelt, unterscheidet solche Beziehungen von anderen inkompatiblen Relationen in *elexiko* (Tab. 1).

Auch wenn sich lexikalische Einheiten, wie es sich aus den Beispielen erkennen lässt, zueinander inkompatibel verhalten, so ist die semantische Inkompatibilitätsbeziehung zwischen ihnen doch anderer Natur.

Die Anzahl und Abfolge der Graduonyme der Paradigmatik von KIND sind nicht unstrittig, sie müssen daher empirisch ermittelt werden. Es werden im Folgenden po-

[14] KWIC (Key Word in Context) ist eine anschauliche Darstellung einer Suchanfrage, die nach unterschiedlichen Kriterien zentriert ausgerichtet und farblich markiert wird.

Stichwort	Inkompatible Partner
1. Mittwoch [Lesart 'Tag']	*Montag, Dienstag, Donnerstag, ...*
2. Auto [Lesart 'Fortbewegungsmittel']	*Bahn, Bus, Flugzeug, Schiff, ...*
3. Kind [Lesart 'sehr junger Mensch']	*Erwachsener, Jugendlicher, Neugeborenes, Säugling*

Tabelle 1: Beispiel für die Relation der Inkompatibilität in *elexiko*

tenzielle Varianten der graduonymischen Paradigmatik von KIND zusammengestellt und analysiert.

a) erste Variante:

Neugeborenes >> Säugling >> Kind >> Jugendlicher >> Erwachsener

Die in *elexiko* als inkompatible Partner von *Kind* angegebenen Wörter sind hier nach dem Steigerungsgrad des Merkmals [+Alter] in einer graduellen Kette angeordnet. Das kontrastierende Verhältnis der inkompatiblen Partner zu *Kind* ist in *elexiko* mit Beispielen belegt. Korpusbelege (1), (2) und (3) sollen den kontrastiven Gebrauch der Wörter veranschaulichen.

(1) Die intensivpflegerische Betreuung von **Neugeborenen, Säuglingen, Kindern und Jugendlichen** ist Schwerpunkt einer Fortbildungstagung. Diese findet bis Samstag in Innsbruck statt. (Tiroler Tageszeitung, 21.09.2000, Kinder benötigen spezielle intensivmedizinische Pflege.)

Da sowohl *Neugeborenes* als auch *Säugling* zum Entwicklungsstadium der Kindheit gehören, subsumiert *Kind* diese Begriffe:

(2) Der Herzfehler soll so schnell als möglich und in nur einem Eingriff korrigiert werden. 55 Prozent der operierten **Kinder** waren **Säuglinge**, 21 Prozent **Neugeborene**. (Salzburger Nachrichten, 28.12.1996, Rasche Korrektur: 114 Kinder in Linz am Herzen operiert.)

Dahingegen weisen die Begriffe *Jugendlicher* und *Erwachsener* höhere Altersstufen auf. Insofern figuriert *Kind* in dieser Gruppe nicht als Hypergraduonym, sondern als inkompatibler Partner (als Graduonym oder graduonymischer Partner):

(3) Drei Musiker brachten mit ihren rhythmischen Klängen Bewegung in den Raum. **Kinder, Jugendliche** und **Erwachsene** wippten und klatschten zur Musik. (St. Galler Tagblatt, 16.09.1997, Alpsaison abgeschlossen.)

Es fehlt also in dieser Reihenfolge ein Oberbegriff für die ganze Gruppe in Form eines Lexems (lexikalisiertes Wort), so dass es alle Wörter der Gruppe (Abb. 3) umfassen könnte. Man könnte die Gruppe als *'Entwicklungsstadium eines Menschen vom Geburt bis zum Erwachsenenalter'* bezeichnen (mehr darüber siehe unten in (c)).

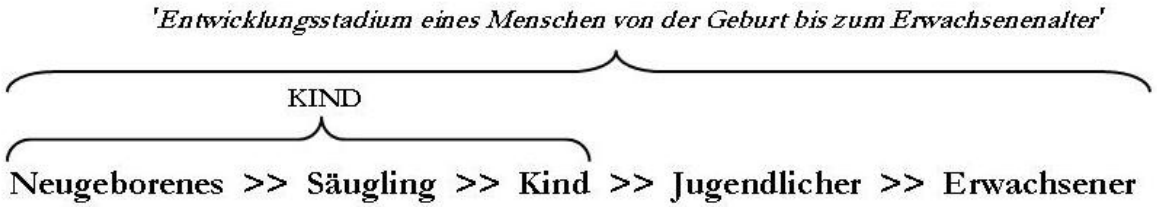

Abbildung 3: Graduonymische Paradigmatik von KIND (erste Variante)

b) zweite Variante:

Neugeborenes >> Säugling >> Kleinkind >> KIND >> Kindergartenkind >> Vorschulkind >> Schulkind

Im Deutschen gibt es mehrere Wörter (hauptsächlich Komposita), wie *Kleinkind, Kindergartenkind, Vorschulkind, Schulkind*[15], die die Lesart 'sehr junger Mensch' haben und daher in die Paradigmatik von KIND eingeschlossen werden können. Aber in *elexiko* ist redaktionell festgelegt, dass Komposita, die das Stichwort als Komponente haben (also Komposita zum Grundwort, wie z.B. *Fahrrad* zu *Rad*, *Schulkind* zu *Kind*), nicht zu den paradigmatischen Relationen der Wörter herangezogen werden. In dieser Arbeit wird dennoch versucht, eine graduonymische Reihenfolge mit dem Hypergraduonym KIND inklusive Komposita zu systematisieren und sie durch Korpusbeispiele zu belegen. Anhand der Sprachdaten der graduonymischen Reihenfolge in (a) wurde ein durch Gradualität gekennzeichneter Bedeutungsgegensatz zwischen den Wörtern *Neugeborenes, Säugling* und *Kind* ermittelt. Das Ziel ist hier zu ermitteln, wie Komposita mit steigerungsrelevantem Merkmal [+Alter], in einer Reihe anzuordnen sind. Es muss zudem erwähnt werden, dass die Anordnung von Lexemen auf Probleme stößt. Erstens ist problematisch, dass es noch Wörter geben könnte, die Bestandteil der graduonymischen Reihe sind. Zweitens fällt es schwer, innerhalb der Skala eine präzise Anordnung bzw. eine stabile Stellung der Lexeme zu gewinnen (vgl. Lyons, 1980, S. 299). Mittels der Korpusanalyse kann aber eine genaue Anordnung der Lexeme, aber auch ihr Verhältnis zueinander erreicht werden. Diesbezüglich ein Beispiel: Bedeutungsunterschiede zwischen *Säugling* und *Kleinkind* werden im typischen Sprachgebrauch innerhalb unterschiedlicher Bedeutungsbeziehungen fokussiert. Korpusbelege zeigen, wie zwischen diesen Wörtern semantisch nahe Beziehungen[16] (1) und kontrastive Bedeutungsbeziehungen (2) bestehen:

(1) Wird bei einem **Säugling** vom Arzt ein mögliches Risiko für SIDS festgestellt, lässt man den Schlaf des **Kleinkindes** mit einem elektronischen Monitor über-

[15] Im Rahmen dieser Arbeit wird die Reihe auf die Wörter *Kleinkind, Kindergartenkind, Vorschulkind, Schulkind* beschränkt. Mit Hilfe der Korpusanalyse könnte die Paradigmatik noch erweitert werden.

[16] Manche Wörterbücher belegen *Säugling* und *Kleinkind* als Synonyme (z.B. Wahrig (2002)). Eine korpusbasierte Untersuchung zur Frage der Sinnrelationen in den Wörterbüchern ist in Storjohann (2005b) ausführlich dargestellt.

wachen, der die Atmung registriert und bei einem Atemstillstand Alarm gibt, damit dem Baby sofort geholfen werden kann. (Kleine Zeitung, 25.09.1996, Hilfe für Säuglinge und Eltern kommt auch ins Haus.)

(2) An vier Tagen, Dienstag und Donnerstag dieser und nächster Woche, versuchen Claudia Kraus und Elke Höpfel vom DRK den Mitarbeitern des Kindergartens in jeweils vier Stunden besondere Erste-Hilfe-Maßnahmen für **Säuglinge, Kleinkinder** und Kinder näher zu bringen. (Mannheimer Morgen, 13.10.1999, "Kinder sind sensibler".)

Die Suchabfrage zum gemeinsamen Vorkommen beider Ausdrücke innerhalb eines Satzes in der passenden Reihenfolge ergab 521 Belege. In 14 Belegen kommen *Säugling* und *Kleinkind* als Synonyme vor. Alle anderen 507 Belege dokumentieren den kontrastiven Gebrauch dieser Ausdrücke. Die häufige Aneinanderreihung der Lexeme macht deutlich, dass ein *Kleinkind* älter als ein *Säugling* ist, also *Kleinkind* auf einer Altersskala rechts vom *Säugling* steht.

Die Bestimmungswörter zum Grundwort *Kind* innerhalb der Komposita *Klein*kind, *Kindergarten*kind, *Vorschul*kind, *Schul*kind signalisieren Altersmerkmale explizit, so dass auf Grund dessen eine Reihenfolge der Wörter festgelegt werden könnte. Welche Verhältnisse in der Sprachverwendung haben *Kindergartenkind* und *Vorschulkind* zueinander und wie steht ihre Relation zu *Schulkind*? Auf den ersten Blick könnte man sich folgende Reihung vorstellen: *Kindergartenkind* >> *Vorschulkind* >> *Schulkind*. Eine Suchanfrage zu *Kindergartenkind* und *Vorschulkind* hat keine eindeutig frequente Belegtreffer zum gemeinsamen Auftreten der Ausdrücke geliefert. Nur in insgesamt 19 Belegen manifestieren sie eine kontrastive Beziehung zueinander.
Es besteht zwischen *Kindergartenkind* und *Vorschulkind* eine Relation zwischen Graduonym und Hypergraduonym (genauso wie die Relation zwischen Hyperonym und Hyponym sowie Grundsynonym und Synonym), so dass jedes *Vorschulkind* (das den Kindergarten besucht) ein *Kindergartenkind* ist, aber nicht jedes *Kindergartenkind* unbedingt ein *Vorschulkind*. Wird bei der Suchabfrage der Abstand bis zu zwei Sätzen erweitert, so zeigen die Suchwörter folgende Verhältnisse in ihrem Gebrauch: kontrastiv – 6mal, bedeutungsnah – 22mal. Die Beispielsätze in (3) und in (4) sollten den kontrastiven und bedeutungsnahen Gebrauch von *Kindergartenkind* und *Vorschulkind* zeigen.

(3) Die jüngsten **Kindergartenkinder** führten einen hübschen Sommertanz auf, das Papageienlied hatten sich die vierjährigen Kinder ausgesucht. Eine rasante Bodenartistiknummer gab es von der Gruppe der fünfjährigen Jungs und Mädchen zu sehen, die **Vorschulkinder** zeigten einen tollen Hip Hop Tanz, der bei den Gästen besonders gut ankam. (Mannheimer Morgen, 16.07.2003, Mäusetanz für Mamas und Papas.)

(4) Förster Thomas Haas führt **Kindergartenkinder** von St. Hildegard durch die Natur Viernheim. **Die Vorschulkinder** der Kindertagesstätte von St. Hildegard

hatten gestern viel Glück mit dem tollen Spätsommerwetter, und so freuten sich alle riesig, als ein zukünftiger Vater sie zu einem Natur-Erlebnistag in den Viernheimer Wald einlud. Mit dem Stadtbus fuhren die Kinder und Erzieherinnen gestern zum Waldschwimmbad, wo im angrenzenden Wald die Tour begann. (Mannheimer Morgen, 16.09.1999, Kleine 'Forscher' erkunden den Wald.)

Auch wenn der kontrastive Gebrauch von diesen Wörtern in den Textsammlungen nicht häufig anzutreffen ist, könnte folgendes Verhältnis bestätigt werden: *Kindergartenkind* >> *Vorschulkind*. Ebenfalls ist die Aneinanderreihung *Kindergartenkind, Vorschulkind* und *Schulkind* nicht belegt, aber es gibt zahlreiche Kontexte, in denen diese drei Wörter inkompatibel gebraucht werden (siehe Beispiel (5)). *Schulkind* trifft häufig in einer einzelnen Kombination (*Kindergartenkind – Schulkind* und *Vorschulkind – Schulkind*) auf und illustriert fast die gleiche Anzahl an Belegen mit beiden Kombinationen (Tab. 2).

Kontrastives Ko-Vorkommen der Wörter	**Abstand innerhalb eines Satzes in gewünschter Reihenfolge**
Vorschulkind – Schulkind	140
Vorschul- und Schulkinder	30
Kindergartenkind – Schulkind	112
Kindergarten- und Schulkinder	54

Tabelle 2: Das 3. Beispiel zu Ko-Vorkommen der Suchwörter

(5) Obwohl das Ferienprogramm in den Gemeinden eigentlich für die **Schulkinder** gedacht ist, hat sich Hannelore Göttlicher - Sportwartin und Übungsleiterin für das Kleinkinder-Turnen des TV 1890 Neckarhausen - diesmal auch eine Attraktion für die **Kindergarten- und Vorschulkinder**, im Alter von drei bei sechs Jahren, einfallen lassen: "Kinder turnen mit den Eltern". (Mannheimer Morgen, 24.08.2000, Flottes Auto rollt durch Turnhalle.)

Als inkompatibler Partner zu *Schulkind* findet sich zudem in den Belegen (innerhalb eines Satzes 36 Belege) (siehe Beispiel (6)) der regional markierte (in der Schweiz) synonymische Partner von *Kindergartenkind – Kindergärtler* (*Kindergärtler* findet sich in Korpora insgesamt 1625mal):

(6) Wenn die **Schulkinder** singen, hören dies die **Kindergärtler**, und wenn diese mit Düften experimentieren, haben auch die Schülerinnen und Schüler etwas zu schnuppern. (St. Galler Tagblatt, 04.07.2000, Schulhaus als "Haus der Sinne".)

Aufgrund der oben angeführten Analyse der Belege kommt eine graduonymische Kette mit dem Hypergraduonym KIND zustande, wie es die Abbildung 4 zeigt.

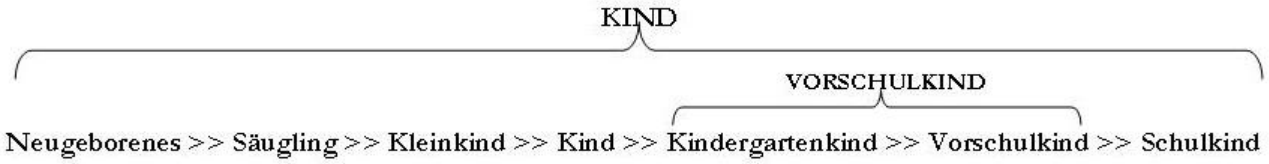

Abbildung 4: Graduonymische Paradigmatik von KIND (zweite Variante)

Selbstverständlich existieren im Deutschen außer den in der Reihe herangezogenen Elementen Wörter, die der graduonymischen Paradigmatik von KIND zugeordnet werden könnten. Die in die Reihe miteinbezogenen Wörter unterscheiden sich durch einleuchtende graduelle Altersmerkmale und werden kontrastierend realisiert, wie es aus den Belegen des tatsächlichen Sprachgebrauchs zu erfassen ist. Andere lexikalische Perspektiven in der Sprache können sich als Zwischenstufen manifestieren.

c) Eine nächste Alternative der graduonymischen Paradigmatik von KIND wäre die vollständige Reihenfolge inklusive Komposita (zur kontrastiven Verwendung der Ausdrücke siehe die Korpusbelege in (1) und (2)):

   Neugeborenes >> Säugling >> Kleinkind >> Kind >> Kindergartenkind
   >> Vorschulkind >> Schulkind >> Jugendlicher >> Erwachsener

(1) Mit Farbe, Papier, Ton und allem was Spaß und Freude bringt, gibt das Team der Kreativität von **Kindergarten- und Schulkindern, Jugendlichen**, Eltern und **Erwachsenen** Raum. (Mannheimer Morgen, 23.09.2004, Kunst für Klein und Groß.)

(2) Eine Sprachberatung für **Vorschul-** oder **Schulkinder, Jugendliche** und **Erwachsene** mit Sprachschwierigkeiten findet im Gesundheitsamt des Main-Taunus-Kreises, Am Kreishaus 1-5 am nächsten und übernächsten Montag von 14.30 bis 16 Uhr statt. (Frankfurter Allgemeine, Tageszeitung, 24.03.2001, Jg. 53.)

Die gesamte Gruppe stellt die drei erkennbaren Altersgruppen vor. Wie oben in (a)) erwähnt, könnte man diese Gruppe unter *'Entwicklungsstadium eines Menschen vom Geburt bis zum Erwachsenenalter'* subsumieren.

Abbildung 5 macht deutlich, dass sich die Graduonymie in diesem System als Skala der aufeinander graduell (inkompatibel) bezogenen Lexeme manifestiert. Eine graduonymische Reihenfolge umfasst i.d.R. stilistisch neutrale lexikalische Ausdrücke. Konnotative und regional begrenzte Varianten jedes Elementes einer graduonymischen Reihe werden in dieses System als synonymische Partner aufgenommen. D.h., jedes Lexem kann in diesem System über seine eigenen synonymischen Partner verfügen. Laut der durchgeführten Kookkurenzanalyse in COSMAS hinsichtlich der Wörter *Säugling* und *Baby* wurde festgestellt, dass sich beide Wörter voneinander stilistisch und durch ihre Verwendungsspezifika unterscheiden. *Säugling* signalisiert dabei eine neutralere Stilebene und Gebrauchsmöglichkeit als *Baby*, so dass beruhend darauf *Säugling* in die

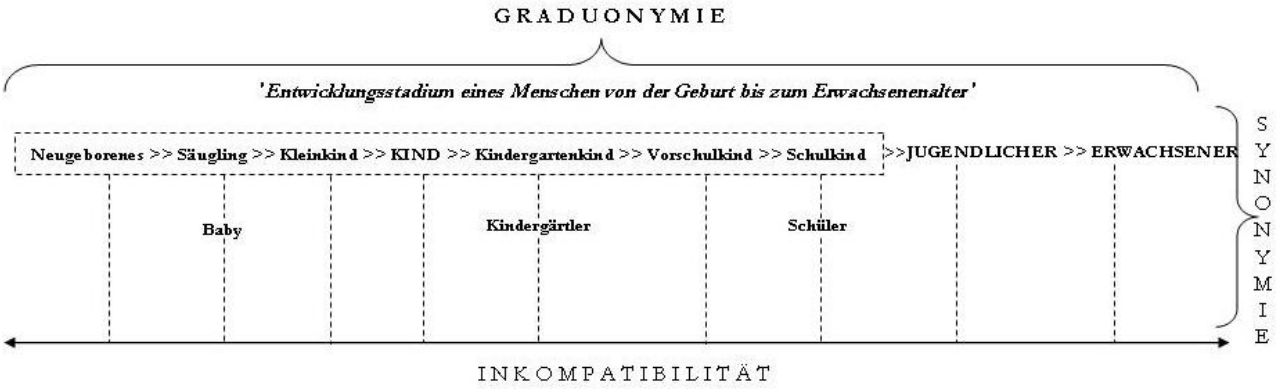

Abbildung 5: Graduonymische Paradigmatik von KIND (dritte Variante)

graduonymische Reihe einbezogen wurde[17]. Aus den Korpusbelegen wurde zu *Kindergartenkind* regional markierter synonymischer Partner *Kindergärtler* extrahiert. Die feinen semantischen Unterschiede in der Verwendung von bedeutungsnahen Wörtern kann unter anderem mit Analyseverfahren der Kontrastierung von Beihnahe-Synonymen (Modul *Contrasting Near-Synonyms*) der Kookkurrenzdatenbank CCDB[18] des Instituts für Deutsche Sprache in Mannheim ermittelt werden. Durch die Erstellung einer kombinierten Merkmalskarte für *Schulkind* und *Schüler* wurde festgestellt, dass beide Wörter häufig Überlappungen in ihrem Gebrauch aufweisen, unterscheiden sich dennoch durch ihre Verwendung in spezifischen thematischen Kontexten: *Schüler* wird als ein für Schulsystem typisches Wort gebraucht (häufige Kookkurenzpartner sind z.B. *Notendurchschnitt, Abitur, schulisch, Lehramt*, etc.); *Schulkind* zeigt deutliche Kontexte auf, die sich auf *Alter* beziehen und *Schulkind* als Kookkurenzpartner mit den Elementen der hier analysierten Paradigmatik von KIND verbinden[19] (z.B. *Kleinkind, Krabbelstube, Hort, Betreuungsangebot, Tagesmutter, Säugling*, etc.).

Aufgrund ihrer Inkompatibilität stehen die Glieder der Skala zueinander in einer Gegensatzrelation. Wenn die graduellen Verschiedenheiten zwischen den Gliedern stark ausgedrückt werden, dann manifestiert sich zwischen den Gliedern der graduonymischen Kette eine Antonymierelation, wie etwa zwischen *Kind* und *Erwachsener* (direkt nacheinander, 242 Belege). In diesem Zusammenhang ist es bemerkenswert, die gewonnenen Korpusdaten mit den Ergebnissen einer anderen Methodik zu vergleichen. Es handelt sich bei dieser Methodik um die durchgeführte Online-Befragung zur Graduonymie des Deutschen[20]. Die Befragung hat das Ziel verfolgt, zu überprüfen, inwie-

[17] Eine detaillierte Analyse zu *Säugling* und *Baby* hat gezeigt, dass sie in von insgesamt 221 Belegen 195mal synonymisch verwendet werden.

[18] Belica, Cyril (2001–2007): Kookkurrenzdatenbank CCDB. Eine korpuslinguistische Denk- und Experimentierplattform für die Erforschung und theoretische Begründung von systemisch-strukturellen Eigenschaften von Kohäsionsrelationen zwischen den Konstituenten des Sprachgebrauchs.

[19] Diese Analysemethode wurde in dieser Arbeit nicht systematisch verwendet, wird aber für die Verifizierung anderer Graduonymiereihungen berücksichtigt.

[20] Das Online-Experiment war auf der Webseite http://www.sfb441.uni-tuebingen.de/~wunsch/nv/release/ aufrufbar.

weit und wie oft die Wörter im Bewusstsein des deutschen Muttersprachlers als Graduonyme wahrgenommen werden. Die graduonymische Paradigmatik von KIND war eine der Reihenfolgen, die von 43 deutschen Probanden bewertet wurden. Bei der gesamten Umfrage standen den Probanden 71 Wortgruppen zur Verfügung. Die Wörter (in alphabetischer Reihenfolge) innerhalb der jeweiligen Wortgruppe sollten nach dem Steigerungsgrad eines spezifischen Merkmals in eine graduonymische Anordnung systematisiert werden. Zur Paradigmatik von KIND wurden acht Wörter, genau die in der Variante (c) angegebenen Wörter außer *Kindergartenkind* in alphabetischer Anordnung, herangezogen: *Erwachsener, Jugendlicher, Kind, Kleinkind, Neugeborenes, Säugling, Schulkind, Vorschulkind.* Die Tabelle 3 soll diese am Beispiel der Paradigmatik von KIND veranschaulichen, wie die Glieder der Gruppe von den Probanden als Graduonyme erkannt und strukturiert wurden.

Variante	**Anordnung der Wörter nach der Beurteilung der Probanden**	**Probanden-zahl**
1	Neugeborenes >> Säugling >> Kleinkind >> Kind >> Vorschulkind >> Schulkind >> Jugendlicher >> Erwachsener	33
2	Neugeborenes >> Säugling >> Kleinkind >> Vorschulkind >> Kind >> Schulkind >> Jugendlicher >> Erwachsener	4
3	Neugeborenes >> Säugling >> Kleinkind >> Vorschulkind >> Schulkind >> Kind >> Jugendlicher >> Erwachsener	1
4	Neugeborenes >> Säugling >> Kleinkind >> Kind >> Schulkind >> Vorschulkind >> Jugendlicher >> Erwachsener	1
5	Säugling >> Kleinkind >> Vorschulkind >> Schulkind >> Kind >> Jugendlicher >> Erwachsener (*Neugeborenes* wurde ignoriert)	1
6	Säugling >> Kleinkind >> Kind >> Vorschulkind >> Schulkind >> Jugendlicher >> Erwachsener (*Neugeborenes* wurde ignoriert)	1
7	Neugeborenes >> Säugling >> Kleinkind >> Vorschulkind >> Schulkind >> Jugendlicher >> Erwachsener (*Kind* wurde ignoriert)	1
8	Neugeborenes >> Kleinkind >> Kind >> Vorschulkind >> Schulkind >> Jugendlicher >> Erwachsener (*Säugling* wurde ignoriert)	1

Tabelle 3: Sprecherurteile für die Paradigmatik von KIND

Wie aus der Tabelle hervorgeht, stimmt die Beurteilung von 33 Probanden (76,7%) in der Variante 1 fast mit den Ergebnissen der Korpusbefunde überein. Bei den Bewertungen von 4 Probanden in der Variante 2 und einem Probanden in der Variante 3 bestehen die Abweichungen bei der Anordnung von *Kind.* Das ist keine wesentliche Verschiedenheit, weil Kind zwischen *Säugling, Kleinkind, Vorschulkind* und *Schulkind* als Hypergraduonym fungiert und über eine abweichende Stellung innerhalb dieser Wörter verfügt. Die Ergebnisse in (2) und (3) sind ebenfalls mit den Korpusdaten kompatibel. In den Beurteilungen in (5; 6; 7; 8) wurde je ein Wort innerhalb der Gruppe ignoriert. Vergleicht man die Ergebnisse der Korpusanalyse mit denen der Sprecherbefragung, so wird deutlich, dass Korpusverfahren bei der Vervollständigung

der graduonymischen Reihen von großer Bedeutung sein können. Die in der Online--Befragung vorhandene lexikalische Lücke zwischen *Kleinkind* und *Vorschulkind* wurde durch die Analyse des tatsächlichen Gebrauchs der Wörter in den Belegsammlungen durch *Kindergartenkind* ergänzt. So kann man aufgrund der Korpusanalyse und der Befragung mit 88 % positiver Bewertung von Probanden die Reihenfolge *Neugeborenes* >> *Säugling* >> *Kleinkind* >> *Kind* >> *Kindergartenkind* >> *Vorschulkind* >> *Schulkind* >> *Jugendlicher* >> *Erwachsener* als stabile graduonymische Reihenfolge annehmen.

7 Fazit

Die Untersuchungen belegen, dass sich aufgrund der Bedeutungsdifferenzierungen zwischen lexikalischen Einheiten lexikalische Gruppen graduonymischer Natur konstituieren lassen. Vermutungen über die Existenz derartiger paradigmatischen Strukturen wurden zunächst durch die eigene Intuition, Analyse von Wörterbüchern, Nutzerbefragungen, und schließlich im Rahmen dieser Arbeit durch die Korpusuntersuchung verifiziert. Die Differenzierung der Graduonymie als eine Art der Sinnrelation erweitert die Paradigmatik von Wörtern. Lexikographische Ressourcen können davon profitieren, die aufgrund eines graduierbaren Merkmals abgestuften Wörter durch eine eigenständige Sinnrelation zu erfassen.

Literatur

Agricola, C. and Agricola, E. (1992). *Wörter und Gegenwörter*. Mannheim [u.a.]: Dudenverlag.

Agricola, E. (1982). Ein Modellwörterbuch lexikalisch-semantischer Strukturen. In *Wortschatzforschung heute. Aktuelle Probleme der Lexikologie und Lexikographie*, pages 9–22. Leipzig: Verlag Enzyklopädie.

Aripzhonova, S. (1994). *Ŏzbek tilida lughaviy graduonimija*. PhD thesis, Toshkent.

Blank, A. (2001). *Einführung in die lexikalische Semantik*. Tübingen: Niemeyer.

Coseriu, E. (1970). *Einführung in die strukturelle Betrachtung des Wortschatzes*. Tübinger Beiträge zur Linguistik 14. Tübingen: Narr.

Coseriu, E. (1978). *Probleme der strukturellen Semantik*. Tübingen: Narr.

Cruse, D. A. (1980). Antonyms and gradable complementaries. In *Perspektiven der lexikalischen Semantik*, pages 14–25. Bonn: Bouvier.

Cruse, D. A. (1986). *Lexical Semantics*. Cambridge: Cambridge University Press.

Cruse, D. A. (2002). Paradigmatic relations of inclusion and identity III: Synonymy. In *Lexikologie/Lexicology. Ein internationales Handbuch zur Natur und Struktur von Wörtern und Wortschätzen*, pages 485–497. Berlin [u.a.]: de Günter.

Cruse, D. A. (2004). *Meaning in Language. An Introduction to Semantics and Pragmatics*. Oxford: Oxford University Press.

Duden (1986). *Duden – Das Bedeutungswörterbuch*, volume Band 10. 2. Mannheim; Wien; Zürich: Bibliographisches Institut.

Duden (2004). Duden - Das Synonymwörterbuch. In: PC–Bibliothek: Nachschlagewerke von Brockhaus, DUDEN, Meyer und Langenscheidt [CD Rom].

Filipec, J. (1966). Probleme des Sprachzentrums und der Sprachperipherie im System des Wortschatzes. *Travaux linguistiques des Prague 2*, pages 257–275.

Görner, H. and Kempcke, G. (2005). *Wörterbuch Synonyme*. Leipzig: Deutscher Taschenbuch Verlag.

Heusinger, S. (2004). *Die Lexik der deutschen Gegenwartssprache*. München: Wilhelm Fink Verlag.

Lemnitzer, L. and Zinsmeister, H. (2006). *Korpuslinguistik*. Tübingen: Gunter Narr.

Lutzeier, P. R. (1981). *Wort und Feld. Wortsemantische Fragestellungen mit besonderer Berücksichtigung des Wortfeldbegriffes*. Tübingen: Niemeyer.

Lutzeier, P. R. (1995). *Lexikologie. Ein Arbeitsbuch*. Tübingen: Stauffenburg.

Lyons, J. (1980). *Semantik. Band I*. München: C.H.Beck.

Ne'matov, H. et al. (1989). Leksik mikrosistema va uning tadqiq metodikasi (sistem leksikologija tezislari). *Ŏzbek tili va adabijoti zhurnali*, (6):35–40.

Ne'matov, H. et al. (1995). *Ŏzbek tili sistem leksikologijasi asoslari*. Toshkent: Ŏqituvchi.

PC-Bibliothek (2001). Nachschlagewerke von Brockhaus, DUDEN, Meyer und Langenscheidt, [CD Rom].

Rachmanov, I. (1983). *Nemecko–russkiĭ sinonimiceskiĭ slovar' /Deutsch-russisches Synonymwörterbuch*. Moskva: Russkiĭ Jazyk.

Schippan, T. (2002). *Lexikologie der deutschen Gegenwartssprache*. Tübingen: Niemeyer.

Storjohann, P. (2005a). Paradigmatische Relationen. In *Grundfragen der elektronischen Lexikographie. elexiko – das Online-Informationssystem zum deutschen Wortschaftz*, pages 249–264. Berlin [u.a.]: de Günter.

Storjohann, P. (2005b). Sinnrelationen in Wörterbüchern - Neue Ansätze und Perspektiven. *ELiSe: Essener Linguistische Skripte elektronisch*, Jahrgang 5(Heft 2):35–61.

Storjohann, P. (2006). Kontextuelle Variabilität synonymer Relationen. *OPAL – Online publizierte Arbeiten zur Linguistik. Mannheim: Institut für Deutsche Sprache*, (1):24 S.

Vokhidova, N. (2007). Lexikalisch-semantische Graduonymie im Deutschen. In Kunze, C., Lemnitzer, L., and Osswald, R., editors, *Proceedings of GLDV-2007 Workshop on Lexical--Semantic and Ontological Resources*, Informatik-Berichte, pages 119–128.

Vokhidova, N. (2009). (im Erscheinen) Ergebnisse einer web-basierten Umfrage zur Graduonymie. In *Akten des 43. Linguistischen Kolloquiums: Pragmantax II. Zum aktuellen Stand der Linguistik und ihrer Teildisziplinen*. Peter Lang.

Wahrig, G. (1986). *Wahrig – Deutsches Wörterbuch.* München: Mosaik–Verlag.

Wahrig, G. (2002). *Wahrig – Synonymwörterbuch.* Gütersloh [u.a.]: Wissen–Media–Verlag.

Wiegand, H. E. (1998). *Wörterbuchforschung: Untersuchungen zur Wörterbuchbenutzung, zur Theorie, Geschichte, Kritik und Automatisierung der Lexikographie.* Berlin und New York: de Gruyter.

Internetressourcen

ELDIT = http://dev.eurac.edu:8081/MakeEldit1/Eldit.html

elexiko = http://www.owid.de/elexiko_/index.html

GermaNet = http://www.sfs.uni-tuebingen.de/GermaNet/

Wortschatz–Portal = http://wortschatz.uni-leipzig.de/

CCDB = http://corpora.ids-mannheim.de/ccdb/

Kiril Simov

Ontology-based Lexicon of Bulgarian

In contrast to morphological and syntactic processing semantic annotation based on domain ontology is still underdeveloped for Bulgarian. On the other hand, the prerequisites for an ontological annotation are already available. These are as follows: a morphosyntactic tagger for Bulgarian with more than 95% accuracy; a dependency parser with more than 84% accuracy; a general chunker and a named entity grammar. Semantic annotation is, therefore, the next logical step. We consider the following to be the minimal set of semantic resources:

- a lexicon for Bulgarian mapped to an upper ontology as a mechanism to cover the common lexicons in domain texts, and mapped to domain ontologies to cover domain terminology;
- an annotation grammar for Bulgarian, based on syntactic knowledge of Bulgarian and conceptual information from the ontology. It comprises grammar rules for recognition of lexical units in the text and rules for selecting the right interpretation in context;
- a corpus, manually annotated with ontology information in order to train the machine learning component for automatic word sense disambiguation — selecting the appropriate concept for a lexical unit in context.

In this paper, we will focus on the description of the lexicon. For that purpose we first present the structure of the domain ontology and then a model of `ontology-to-text` relations which facilitates the text annotation. The ontology-based lexicon for Bulgarian is viewed as part of the model. It is constructed in an incremental manner.

1 Introduction

There exist various types of lexicons. Some of them just register the lexemes and/or their wordforms, while others reflect valency or lexical relations, such as synonymy, antonymy, etc. There are also the so-called thesauri lexicons, which combine the formal and lexical aspects of the lexemes. The most popular thesaurus is the WordNet lexical database. On top of these types of lexicons there exist generative ones (in Pustejovsky's sense), which connect the words in the lexicon to some kind of linguistic ontology (e.g. EuroWordNet, SIMPLE lexicon). For the purposes of the semantic annotation of domain text another type of lexicon is needed, namely a lexicon mapped to a domain ontology. However, such a lexicon cannot exist in isolation from the more general lexicon

which covers non-domain words. The combination of the general and domain-specific lexicons is very important for the semantic annotation on domain level since the usage of the domain terms in the text depends on the semantics of the common words in the context. Semantic annotation is an enhancement in the area of language resources after the creation of morphologically and syntactically annotated corpora. The importance of semantic annotation became a hot topic within the Semantic Web initiative. Although much work is already done in this area, the term 'semantic annotation' is not yet well defined - see (Kiryakov et. al, 2005) and citations therein. In our work we think of a text as consisting of two types of (non-linguistic) information: (1) ontological classes and relations, and (2) world facts. The ontological part determines generally the topic and the domain of the text. The corresponding 'minimal' part of ontology implied by the text will be called 'ontology of the text'. The world facts represent an instantiation of the ontology in the text (here entities like beliefs, claims, attitudes, etc. are also included). Both types of information are called uniformly 'semantic content of the text'. Both components of the semantic content are connected to the syntactic structure of the text. Any (partial) explication of the semantic content of a text will be called semantic annotation of the text. Defined in this way, the semantic annotation could contain also some pragmatic information and actual world knowledge.

To support the semantic annotation we rely on an ontology-based lexicon. We assume that there is a domain ontology which is used in the process of annotation. A domain ontology comprises three layers: domain, middle and upper. The lexicon is mapped to the ontology. This mapping is based on relations between the meaning of the lexical units in the lexicon and concepts (relations and instances) in the ontology. Thus, we assume that the ontology contains the conceptual information necessary to model the word senses in the lexicon. The advantage of using an ontology is that the reflections of the conceptualization of the world become explicit.

The motivation for the construction of such a lexicon is the need for more precise semantic annotation. In order to ensure this, the lexicon has to provide more complex conceptual information than the one in computational lexicons like WordNet. The second requirement for the ontology-based lexicon is the coverage of the words in the text. The lexicon has to cover not only the domain terms, but also the non-specialized language. This is necessary for ensuring enough explicit knowledge for the application of word sense disambiguation methods based on statistics. Since the development of a general ontology to support all the lexical units in a language is an intractable problem, we construct the ontology in an incremental way from the upper ontology to the middle and domain specific ontology. Then lexical units are mapped to this ontology via two relations — `equality` and `subsumption`. The first is used when the appropriate concept for a meaning of some word is already represented in the ontology. The latter is used when such a concept is missing and only super-concepts are available.

We place a special emphasis in this paper on the role of metonymy and regular polysemy. They are encoded as specific patterns extracted from a semantically annotated corpus and reflect the conceptual structure of the ontology. The lexicon is also connected to an annotation grammar which establishes a relation between the ontology and

the text. Other phenomena like metaphoric relations and near-synonyms are not treated at the moment in this version of the lexicon, but the model provides possibilities for extensions to represent them in future.

Although being part of the annotation process, the concept annotation grammar remains beyond the scope of this paper. When mentioned, it is only with the aim to highlight the lexicon within a context. More specifically, our idea is the following: the ontology delivers the concepts within the world, the lexicon stores the wordings of those concepts (both lexicalized and non-lexicalized), and the text disambiguates the concepts within a concrete discourse pattern.

The structure of the paper is as follows: the next section discusses related works; then the architecture of a domain ontology comprising an upper ontology, a middle ontology, domain specific part and linguistic knowledge mapped to them is explained; the fourth section presents the creation of the ontology-based lexicon of Bulgarian. It first discusses a model for domain lexicons used for semantic annotation. Then the model is extended to cover the general language lexicon. At the end the encoding of some special phenomena is presented; and the last section concludes the paper.

2 Related Works on Ontology and Lexicon

Ontologies and lexicons are artifacts reflecting the human abilities for representing, processing and managing linguistic and conceptual knowledge. As such, they allow for the combination of many different approaches. A recent overview of the relation between ontologies and lexicons is presented in (Hirst, 2004). The paper discusses the structure of lexical entries, the knowledge recorded in them and mechanisms for interrelation of the lexicon elements. Special attention is given to the definition of 'word sense', its conceptual structure, relations between senses and problematic cases. The main topics under discussion are near-synonyms, gaps in the lexicon, and linguistic categorizations that are not ontological. We treat these topics as follows.

First, we assume that the lexicon is based on the ontology, i.e. the word senses are represented by concepts, relations or instances. Near-synonyms are words that share the same central conceptual information, but differ in the additional information they provide to the semantic interpretation module, such as small changes in the denotation, different implications, speaker attitude, etc. Our model does not solve this problem completely. In it only the central part of the meaning of a word that can be represented. The additional parts of the meaning (context related variations) can be encoded as additional information in the lexical entry or as an extension of the ontology where it is appropriate — similarly to the model used in (Edmonds and Hirst, 2002). The problem of lexical gaps is solved by allowing the storage of free phrases. Similarly, gaps in the ontology (a missing concept for a word sense, for example) are solved by appropriate extensions of the ontology. Linguistic categorizations that are not ontological are not treated in our model.

As it was mentioned above, the construction of a Bulgarian ontology-based lexicon is motivated by the need to introduce more world knowledge into the semantic analysis

of texts. (Morris and Hirst, 2004) points out that most of the lexical relations necessary to determine the semantic content of lexical units are non-classical in contrast to the classical ones, i.e. `hyponymy`, `meronymy`, `antonymy`. The non-classical relations are specific to some classes of meanings, i.e. `made-of`, `used-for`, etc. In our case we assume that these relations are represented in the ontology. Thus, they are formally defined, can be used for the purposes of semantic inference and can be used for the representation of some language phenomena like polysemy, metonymy, etc.

Regarding the complexity and precision of a given ontology we follow the definition in (Guarino, 2000). It represents the following classification of ontologies:

- **Lexicon**: *Machine Readable Dictionaries; Vocabulary with natural language definitions*
- **Simple Taxonomy**: *Classifications*
- **Thesaurus**: *WordNet; Taxonomy plus related-terms*
- **Relational Model**: *Light-weight ontologies; Unconstrained use of arbitrary relations*
- **Fully Axiomatized Theory**: *Heavy-weight ontologies.*

The classification starts with a less formal and knowledge-poor ontology (hence — simple lexicons) and ends with heavily constrained theories about the world. Sometimes the first three elements of the classification are not considered as ontologies, because the ontological information is represented mainly implicitly. As it was pointed out to us by one of the reviewers this hierarchy shows the transition from lexicon to ontology. In our view such a transition supports the mapping between the ontology and lexicon. Our attempt is to move the current semantic lexicons for Bulgarian from the level of thesaurus to the level of light-ontologies (as a minimum).

Our approach draws in many respects on the work done on WordNet (Fellbaum, 1998), EuroWordNet (Vossen, 1998), SIMPLE (Lenci et. al, 2000). With WordNet-like lexicons — (Fellbaum, 1998) and (Vossen, 1998) — we share the idea of grouping lexical units around a common meaning and in this respect the term groups in our model correspond to synsets in the WordNet model. The difference is that the meaning is defined independently in the ontology. With the SIMPLE model (Lenci et. al, 2000) we share the idea to define the meaning of lexical units by means of an ontology, but we differ in the selection of the ontology which in our case represents the domain of interest, and in the case of SIMPLE reflects the lexicon model: Generative Lexicon — (Pustejovsky, 1995). Similar is the connection with EuroWordNet.

With the LingInfo model — (Buitelaar et. al, 2006a), (Buitelaar et. al, 2006b) and (Romanelli et. al, 2007) — we share the idea that grammatical and context information also needs to be presented and linked to the ontology, but we differ in the implementation of the model and the degree of realization of the concrete language resources and tools.

Finally, we would like to mention the work within the Ontology Semantics (Nirenburg and Raskin, 2004) which is very similar to our model except that we use existing ontologies like DOLCE and we allow for an incremental construction of the lexicon.

3 The Structure of a Domain Ontology

Our work is based on a model developed for the annotation of domain concepts in a text. In this model we assume that the ontology is the starting point for the creation of the `ontology-to-text` relation. The structure of a domain ontology can be defined (at least) with respect to: (1) the collection of concepts represented in the ontology; and (2) the complexity of the concept definitions.

Independently from the methodology for ontology creation, the concepts represented in the ontology can be distributed on the following layers which reflect the generality of the conceptual information:

- **Domain layer**.

  At this layer we have the domain concepts and relations representing the main notions in the domain. These concepts and relations are used for solving different tasks, such as the representation of domain knowledge, the representation of common conceptualization for information exchange in the domain, the semantic annotation of domain texts, etc.

- **Upper layer.**

  The alignment of the domain layer to an upper ontology is an obligatory step in each ontology creation methodology. This alignment ensures several properties of the domain ontology: (1) its consistency with the design of the upper ontology; (2) inheritance of the knowledge represented in the upper ontology. The inheritance requires the imposition of more specific constraints reflecting the structure of the domain.

- **Middle layer.**

  This layer contains concepts and relations which are neither part of the upper layer, nor of the domain one, but play an important role for the alignment between them. For example, 'carpet' is in the domain layer for the Home Textile ontology and 'artifact' is in the upper layer, but the concept for 'covering' which is more specific than 'artifact' and more general than 'carpet' (defined as textile floor covering) is in the middle layer. This layer is the result of the ontology creation practice and depends on the coverage of the domain and the range of concepts defined in the upper ontology. In our view it is a useful instrument for transition from the domain to the upper layer.

An additional layer related to the conceptual information is the linguistic information represented by a lexicon and a grammar. This information is necessary in all cases where

the ontology interacts with natural language, for example in the analysis of texts, when navigating the ontology, in ontology based searches, etc.

- **Language layer.**

  It is assumed that the ontology with its three layers is language independent, formalized in some ontology representation language. In practice, such an ontology would be incomprehensible to humans and therefore has to be mapped to some linguistic resources. This mapping is required for at least two reasons: to allow to present the ontology to users who are not ontology engineers, and to support semantic analysis and retrieval of texts. Thus, as a minimum it is necessary to have a lexicon mapped to the concepts and the relations in the ontology. For example, the concept 'scanner' would have several possible wordings in English, such as 'scanner', 'image scanner', 'digital scanner'.

The following figure shows the structure of a domain ontology:

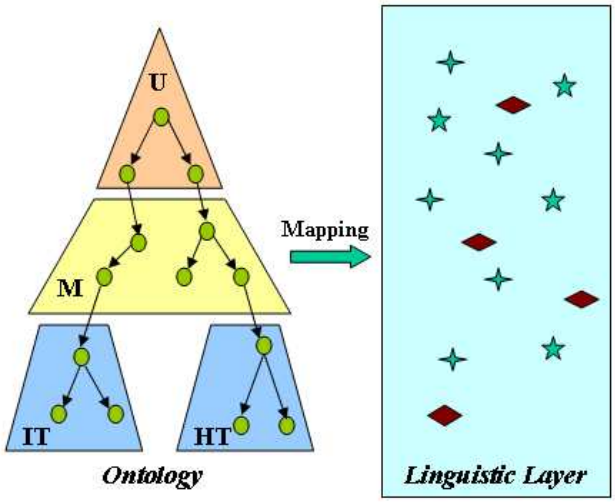

Fig 1. Domain ontology structure.

Here on the left side we have two domain ontologies (IT for the domain of Information Technology for End Users and HT for the domain of Home Textile) aligned to the middle layer and the upper layer. The linguistic layer consists of lexical units, grammar rules, disambiguation rules, etc. The mapping between the ontology and the linguistic layer is the way to define the `ontology-to-text` relation. This relation supports the semantic annotation of text. It generally comprises a lexicon and a grammar.

We have used this structure of the ontology in three European projects - LT4eL, AsIsKnown and LTfLL. In each of them we have used as an upper ontology the DOLCE Ontology (Masolo et. al, 2002) for several reasons: (1) it is constructed on a rigorous basis which reflects the OntoClean methodology (Guarino and Welty, 2002); (2) it contains many useful relations and axioms which constrain the interpretation of the ontology; (3) it is represented in OWL-DL. For the middle layer we have used OntoWordNet (Gangemi et. al, 2003) - a version of WordNet aligned to DOLCE. The domain layer is created for each domain. The result of the three layers is a domain ontology with a better structuring of the concepts and relations. In addition, relations and axioms are inherited from the DOLCE upper part to the specific domain layer. The linguistic layer was implemented via domain lexicons, presented in the next section, and concept annotation grammar, described in (Simov and Osenova, 2007) and (Simov and Osenova, 2008).

4 An Ontology-based Lexicon of Bulgarian

In this section we present our work on a Bulgarian ontology-based lexicon. First we describe the structure of domain lexicons developed for the projects mentioned above. Then we extend their structure in order to overcome the problems we faced with respect to the annotation of domain texts.

4.1 Domain Lexicons

In order to support semantic annotation of domain texts we have defined an `ontology-to-text` relation based on three elements — domain ontology, domain lexicon and concept annotation grammar. The model of the domain lexicon is based on the assumption that the ontology has a central role in the definition of the `ontology-to-text` relation and the language information reflects the available conceptual information in the ontology. The mapping is directed from the ontology to the lexicon, then from the lexicon to the grammar and then to the text. For each concept (relation, instance) in the ontology the lexicon contains at least one lexical unit. This requires the lexicon to contain non-lexicalized (fully compositional or free) phrases as well[1]. Availability of different lexical units (lexicalized or not) for a given concept is used as a basis for the construction of the annotation grammar. This availability allows us to capture different wordings of the same meaning in a text. In general, a concept might have a few terms connected to it and a (potentially) unlimited number of free phrases expressing this concept in the language. Some of the free phrases receive their meaning compositionally regardless of their usage in a given text, other free phrases denote the corresponding concept only in a particular context. In our lexicons we register as many free phrases as possible in order to have better recall on

[1] The presence of free phrases in the lexicon is also motivated by the fact that lexicalization is not a discrete feature. There are many different degrees of lexicalization. Thus the free phrases are one extreme end of the scale.

the semantic annotation task. In cases when a lexicalized concept is missing in the ontology we modify it. The model was used for the construction of lexicons in several languages for two domains — Information Technology and Home Textiles. When we have lexicons in several languages mapped to the same ontology we ensure a certain level of multilinguality. The following figure shows the mapping from the ontology to the lexicon:

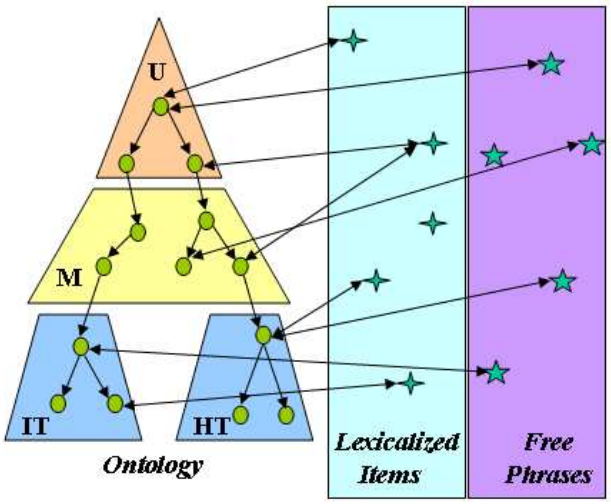

Fig 2. Ontology to lexicon mapping using `equality` relation.

The lexicon plays a double role within the three projects. First, it interrelates the concepts in the ontology to the lexical knowledge used by the grammar in order to recognize the role of the concepts in the text. Second, the lexicon represents the main interface between the user and the ontology. This interface allows for the ontology to be navigated or represented in a way natural for the user. For example, the concepts and relations might be named with lexical units employed by the users in their everyday activities and in their own natural language (e.g. Bulgarian). This could be considered as a first step to a contextualized usage of the ontology in the sense that the ontology could be viewed through different terms depending on the context. For example, the colour names will vary from very specific terms within the domain of carpet production to more common words used when the same carpet is part of an interior design. Thus, the lexical entries contain the following information: lexical units (words or phrases), contextual information determining the context of the their usage, grammatical features determining their syntactic realization within the text. In the current implementation of the lexicons the contextual information is simplified to a list of a few types of user

roles. In the home textile domain the users are producer, retailer, interior designer, etc. In the eLearning domain they are student, teacher, tutor, administrator. Depending on the role, the ontology is filtered with the help of the lexicon, which focuses preferably on the user-specific words. For example, the carpet producer would be interested in the number expressions of the colours as presented in a standard, the various types of carpets and the detailed structure of the carpet. On the other hand, the interior designer would like to explore the relation of the carpet to other parts of the interior (wall, paintings, furniture, etc.).

We could summarize the connection between the ontology and the lexicons in the following way: the ontology represents the semantic knowledge in the form of concepts and relations with appropriate axioms; and the lexicons represent the ways in which these concepts can be realized in texts in the corresponding languages. Of course, the ways in which a concept could be represented in the text are potentially infinite in number, thus, we aimed at representing in our lexicons only the most frequent and important words and phrases. Here is an example of an entry:

```
<entry id="entry-34">
 <owl:Class rdf:about="http://www.asisknown.org/AIKHT#Carpet">
   <rdfs:comment>a piece of thick heavy fabric
                 used to cover a floor</rdfs:comment>
   <rdfs:subClassOf>
     <owl:Class rdf:about=http://www.asisknown.org/AIKHT#FloorCovering/>
   </rdfs:subClassOf>
 </owl:Class>
 <def>a piece of thick heavy fabric used to cover a floor</def>
 <termg lang="en">
   <term shead="1">carpet</term>
   <term>carpeting</term>
   <term>rug</term>
   <term type="nonlex">textile floor covering</term>
   <def>a piece of thick heavy fabric used to cover a floor</def>
   <gramline>reference to finite state grammar</gramline>
 </termg>
</entry>
```

Each entry in the lexicons contains the following types of information: (1) information about the concept from the ontology which represents the meaning for the terms in the entry; (2) explanation of the concept meaning in English; (3) a set of lexical units (in domain lexicon we call them terms) in a given language representing the concept; and (4) relation to grammar rules. The concept part of the entry provides the minimum information necessary for a formal definition of the concept. The English explanation of the concept meaning facilitates human understanding. The set of terms stands for different wordings of the concept in the respective language. One of the terms is a representative for the term set. Note that this is a somewhat arbitrary decision, which

might depend on the frequency of term's usage or on the expert's intuition. This representative term will be used where only one of terms from the set is called for, for example as an item of a menu. In the example above we present the set of English terms for the concept 'carpet'. One of the terms is non-lexicalized — the attribute `type` with value **"nonlex"**. The first term is representative for the term set and it is marked-up with the attribute `shead` with the value **"1"**. The elements `gramline` provide links to linguistic features of the terms like lemmatized variants of the terms, implementation as regular expressions to be compiled as finite state automata.

The second component of the `ontology-to-text` relation, namely the concept annotation grammar, is ideally considered to be an extension of a general deep grammar of a given language which is adopted in the concept annotation task. Minimally, the concept annotation grammar consists of a chunk grammar for concept annotation and (sense) disambiguation rules. The chunk grammar for each term in the lexicon contains at least one grammar rule for the recognition of the term. The annotation with grammatical features and the lemmatization of the text are considered a preprocessing step. The disambiguation rules exploit the local context in terms of grammatical features, semantic annotation and syntactic structure, and also the global context such as the topic of the text, the discourse segmentation, etc. Currently we have implemented chunk grammars for several languages. We have implemented a very simple disambiguator which uses an unigram model. The annotation grammar is implemented within the CLaRK System (Simov et. al, 2001).

The relation `ontology-to-text` implemented in this way provides facilities for solving different tasks, such as ontology search (including crosslinguistic search), ontology browsing, ontology learning. In order to support multilingual access to a semantically annotated corpus we have to implement the relation for several languages using the same ontology as a starting point. In this way we implement a mapping between the lexicons in these languages and also a comparable annotation of texts.

4.2 The Extended Model

The main problem with the model of the `ontology-to-text` relation, described in the previous section is the fact that the lexicon is mapped only in its domain part to the ontology. Thus, the annotation of domain texts with domain concepts is very sparse. For example, in the IT domain we have annotated 8 concepts within 100 tokens (with 14.8 tokens per sentence = 1.19 concepts per sentence at average). This sparse annotation blocks possibilities for using better methods for word sense disambiguation. This holds when the lexical units in the domain lexicon are ambiguous among themselves or with respect to the lexical units from the general lexicons. For example, the concepts 'key-of-keyboard', 'key-of-database' and 'key-for-door' have the same wording in English ("key") and the last one is not from the domain ontology. Therefore, we need a much better semantic annotation than one which just uses the domain terms and grammar constructed on their basis.

To achieve such a better annotation we consider two tasks to be solved: (1) to ensure better coverage of the text with conceptual information and (2) to exploit a better disambiguation model using this information. For the first task we envisage two interconnected solutions: (1) to improve the annotation grammar, and (2) to provide an interaction with the general lexicon. The second task is not discussed here.

The annotation grammar can be improved in several directions. Most prominent seems to be the usage of the syntactic structure and co-referential relations in order to distribute the domain knowledge to general lexical units or phrases in the text. For example, if a document is about 'desktop publishing system' very often this concept can be referred with expressions like 'publishing system' or 'system'. Similarly, predicates impose semantic restrictions on their arguments. The interaction with the general lexicon can be achieved via connecting of the domain lexicon to lexicons like WordNet. The interaction between the two solutions is as follows. First, the words in the text receive their annotation from the domain and the general lexicons. The grammar ensures additional distribution of the domain annotations via the co-referential relations and syntactic structures. Finally the disambiguation module selects the right annotations. In order to support this processing the domain terms and the common words in the context have to share the semantic annotation. In order to ensure this we have to augment the general lexicon with appropriate semantic information.

Ideally, each meaning of a lexical unit from the general lexicon has to be present in the ontology in order to use the model of `ontology-to-text` relation discussed above. Unfortunately, such an ontology does not exist yet. Thus, we have to use a smaller ontology and to change the implementation of the `ontology-to-text` relation.

On the basis of the gained experience within the projects mentioned above we conclude that there exists a relatively stable upper and middle part for each of the domain ontologies. Therefore, we think that a first step for the creation of an appropriate lexical resource for semantic annotation is the building of an upper-middle layer ontology. This step can provide the necessary semantic information for the tasks of word sense disambiguation. Such an ontology can be used in several ways: (1) for the representation of the general meaning of lexical units in a language; (2) as the basis for the construction of domain ontologies and lexicons; (3) to supply labels in the ontology comprehensible to speakers of human languages. For the concrete implementation of the Bulgarian ontology-based lexicon we use the ontology which results from extension of DOLCE and the upper part of OntoWordNet. This ontology was already used in the construction of the domain ontologies.

In the previous model we have used only the `equality` relation between the conceptual information in the ontology and the meaning of the corresponding lexical units. In this new lexicon this will not be possible because there will be insufficient concepts in the ontology. Thus, we separate completely the lexical information from conceptual one. In the lexicon each lexical entry contains linguistic information just for one lexical unit (representing one meaning). The linguistic information includes morphological and syntactic information. The conceptual information is presented via the relations `equality` or `subsumption`. When there is a concept in the ontology which is equivalent

to the meaning of the lexical unit, then only the relation `equality` is used to connect the lexical unit to the corresponding concept. If there is no equivalent concept for the lexical unit, the relation `subsumption` is used. In this case the lexical unit can be mapped to more than one concept from the ontology, because in the ontology multiple inheritance is allowed. When a lexical unit is mapped to several concepts, its ontology representation is a disjunction of these concepts. The requirement for the mapping via the `subsumption` relation is that the concepts used are the most specific ones available. The following figure depicts the addition of the new relation for the mapping between ontology and lexicon:

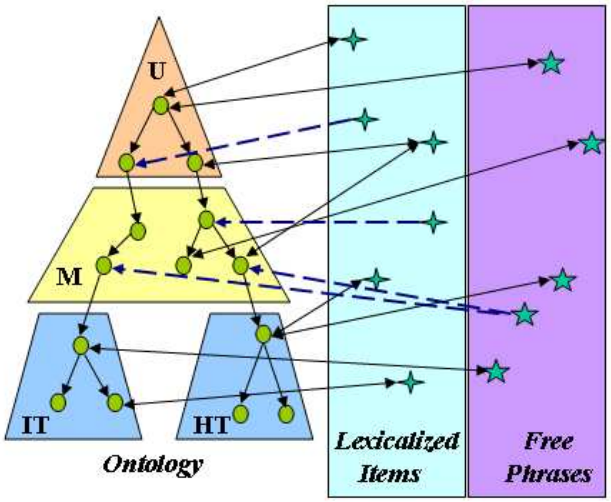

Fig 3. Ontology to lexicon mapping using `equality` and `subsumption` relations.

The dashed arrow represents the `subsumption` relation and the solid double arrow represents the `equality` relation. Concepts from the upper-middle layer are connected to two sets of lexical units — lexical units having a meaning represented by these concepts and lexical units having more specific meaning. The concepts from the domain layer are connected only to lexical units with equivalent meaning. Adding new domain ontologies will improve the precision of the mapping between the ontology and the lexicon.

The actual lexicon is under construction. It is based on several machine-readable dictionaries: a Morphological Dictionary, a Valence Dictionary and an Explanatory Dictionary of Bulgarian. The selection of the lexical units is done on the basis of constructing the lexicon aligned to the upper and middle parts of the ontology where

we encoded about 3000 lexical entries. The rest of the lexical units are selected on the basis of their ranking in a large Bulgarian reference corpus (72 million running words from the BulTreeBank text archive). The ranks are calculated on the basis of an automatic morphosyntactic analysis of the corpus and then lemmatization. For each lemma we consider the frequency in the corpus and in how many documents the lemma has occurred. The lexicon also contains the lexical entries for the two domain ontologies.

4.3 Encoding of Special Phenomena

Including of a formal ontology in the lexicon construction provides many possibilities for using the knowledge, represented in the ontology and the services related to it, such as inference. In this section we present the encoding of some important phenomena for the task of word sense disambiguation: metonymy and verb frames representation. The metonymy covers also a substantial part of the cases of regular polysemy. For an overview on regular polysemy, its representation and importance of this representation see (Barque and Chaumartin, 2009). We assume that the patterns described by the authors can be represented as inference patterns in our model of lexicon to ontology mapping.

A general assumption in the treatment of the above mentioned phenomena is that the related word senses are already represented in the ontology. In this way, the lexical representation of the corresponding patterns (metonymical or frame) is done via appropriate mappings to corresponding concepts in the ontology. The application of such patterns for creation of new senses is not explored in this work.

Let us consider the case of metonymy in more detail. In general, metonymy is defined as a trope in which one entity is used to stand for another associated entity[2]. Our treatment follows the ideas of (Hobbs, 2003) who interpreted the metonymy by introduction of a function which relates the mentioned object with the intended one. The function is different for different cases of metonymy and it can be context dependent. In order to implement the same idea we assume that the function is determined by an inference over the ontology and the context. This function is a composition of relations from the ontology. We consider the representation of such compositions in the lexicon as an important device for facilitation of text annotation. Our view of these compositions is that they are very specific inference rules. In future we will investigate the possibility to encode the metonymy relations reported in the literature (like the ones presented in (Barque and Chaumartin, 2009)) as such special inference rules. Here we present the interpretation of two cases of metonymy.

Let us suppose that we have to annotate the sentence "She was wearing stripe." First we annotate 'stripe' as a kind of `property` and as such it is connected to 'cloth' via the `property-of` relation and 'cloth' is annotated as `material` and it is connected to 'clothing' via the `made-of` relation. The concept 'clothing' is of the relevant type for the object of the verb 'to wear'. Thus, the understanding of the sentence is something like: "She was wearing a clothing made from a textile with a stripe design." The composition

[2] `http://www.sil.org/linguistics/GlossaryOfLinguisticTerms/WhatIsMetonymy.htm`

of the corresponding relations is stored in the lexical entries for the corresponding lexical units. In the case of metonymy this is a better option, because the possible patterns are (potentially) infinite in number. Representing each metonymy usage as a separate meaning will result in many strange meanings for the lexical units. In this way we represent the most frequent metonymy uses as inference patterns and the actual inference is done during the analysis of the discourse where the lexical unit is used metonymically.

When the regular polysemy is an example of metonymy, we represent it in the same way. The different meanings are represented in the ontology as different concepts and these concepts are connected via appropriate relations. The main difference here is that for each of the meanings we construct a separate lexical entry. This means that during the analysis of the text we have to disambiguate between these senses. In some cases more that one of the senses is visible via one usage of the lexical unit. For example, in the sentence "This large book is very interesting." the word 'book' is used simultaneously as a `physical object` selected by 'large' and as an `information object` selected by 'interesting'.

The encoding of verbs is also very important for the task of semantic annotation. We assume that the appropriate information is represented in two ways: (1) in the ontology each verb is connected to an event concept related to the meaning of the verb. In the ontology all the participants (irrespective of whether they are considered to be arguments, adjuncts, etc.) are represented as such via appropriate relations; (2) the linguistic behavior is encoded in the lexicon as a set of frames. These frames determine the role of each participant in the given event. During the annotation the verbs are annotated with the frames from the lexicon and the corresponding relations are connected with appropriate phrases from the text. Some of them are left unconnected when the corresponding participant is not explicitly mentioned in the text.

Currently, we do not represent in the lexicon the relation between the literal meaning of a given word and its metaphorical meaning. In contrast to metonymy, metaphorical meanings are not always closely related in the ontology. They require a special kind of inference by analogy which differs in many respects from the inference necessary to deal with metonymy.

5 Conclusion

In this paper we presented the construction of an ontology-based lexicon for Bulgarian. This lexicon originates from the practical task of semantic annotation of domain texts. Our starting point was the mapping from domain ontologies to terminological lexicons. Due to the sparseness of the resulting concept annotation, the coverage was extended to the general lexicon. One suggestion we make within the model is that merging the upper part of the ontology with the WordNet middle layer would result in a reduced resource which, however, is more understandable to common users. In order to have a better coverage, we rely on two relations between lexical units and the concepts in the ontology: `equality` and `subsumption`. The first is used primarily for the domain

ontology and the second for the middle and upper part of the ontology. Additionally, we encode metonymy and verb frames in the lexicon in order to support a better text annotation.

Our future goals are to implement a system for automatic word sense disambiguation and for detection of metonymical uses in the text. The extension of the lexicon coverage is also one of our tasks. In addition, the general lexicon together with the ontology could be used for the creation of domain ontologies and lexicons. We also plan an annotation of a corpus with concepts from the middle and upper part of the ontology.

6 Acknowledgements

This work has been supported by three European projects: LT4eL (Language Technology for eLearning) (FP6-027391), AsIsKnown (A Semantic-Based Knowledge Flow System for the European Home Textiles Industry) (FP6-028044) and LTfLL (Language Technologies for LifeLong Learning) (FP7-212578).

I would like to thank Petya Osenova for the comments and the discussions on the earlier versions of the paper and to the two anonymous reviewers for their valuable comments and suggestions.

Literatur

Barque L. and Chaumartin Fr. R. (2009). *Regular polysemy in WordNet.* In this volume.

Buitelaar P., Declerck Th., Frank An., Racioppa St., Kiesel M., Sintek M., Engel R., Romanelli M., Sonntag D., Loos B., Micelli V., Porzel R., Cimiano Ph. (2006). LingInfo: Design and Applications of a Model for the Integration of Linguistic Information in Ontologies. In: *Proceedings of OntoLex06, a Workshop at LREC*, Genoa, Italy.

Buitelaar, P., Sintek, M., and Kiesel, M. (2006). A Lexicon Model for Multilingual/Multimedia Ontologies In: *Proceedings of the 3rd European Semantic Web Conference (ESWC06)*, Budva, Montenegro.

Edmonds Ph. and Hirst Gr. (2002). *Near-synonymy and lexical choice.* Computational Linguistics, Vol. 28:2, pp. 105–144.

Fellbaum Chr. (1998). Editor. *WORDNET: an electronic lexical database.* MIT Press.

Gangemi, A., Navigli, R., and Velardi, P. (2003). The OntoWordNet Project: extension and axiomatization of conceptual relations in WordNet. In: *Meersman R, et al. (eds.), Proceedings of ODBASE03 Conference*, Springer.

Guarino N. (2000). *Invited Mini-course on Ontological Analysis and Ontology Design.* First Workshop on Ontologies and lexical Knowledge Bases - OntoLex 2000. Sozopol, Bulgaria.

Guarino, N., and Welty, C. (2002). *Evaluating Ontological Decisions with OntoClean.* Communications of the ACM, 45(2): 61-65.

Hirst Gr. (2004). *Ontology and the lexicon.* In: Steffen Staab and Rudi Studer (editors), Handbook on Ontologies. Springer Verlag, Berlin, Germany. pp 209–229. `http://ftp.cs.toronto.edu/pub/gh/Hirst-Ontol-2003.pdf`

Hobbs J. R. (2003) *Discourse and Inference.* University of Southern California, Marina del Rey, California. Unpublished manuscript. `http://www.isi.edu/~hobbs/disinf-tc.html`

Kiryakov At., Popov B., Terziev Iv., Manov D., and Ognyanoff D. (2005). *Semantic Annotation, Indexing, and Retrieval.* Elsevier's Journal of Web Sematics, Vol. 2, Issue 1.

Lenci A., Busa F., Ruimy N., Gola El., Monachini M., Calzolari N., Zampolli A., Guimier E., Recourcé G., Humphreys L., von Rekovsky U., Ogonowski A., McCauley Cl., Peters W., Peters Iv., Gaizauskas R., Villegas M. *SIMPLE Work Package 2 - Linguistic Specifications, Deliverable D2.1.* ILC-CNR, Pisa, Italy.

Masolo, C., Borgo, S., Gangemi, A., Guarino, N., and Oltramari, A. (2002). *Ontology Library (final).* WonderWeb Deliverable D18, December 2003. http://www.loa-cnr.it/Publications.html.

Morris J. and Graeme Hirst Gr. (2004). Non-Classical Lexical Semantic Relations. In: *Proceedings of the HLT Workshop on Computational Lexical Semantics.* Boston, Massachusetts, USA. pp 46–51.

Nirenburg S. and Raskin V. (2004). *Ontological Semantics.* MIT Press.

Pustejovsky J. (1995). *The Generative Lexicon.* MIT Press. Cambridge, MA, USA.

Romanelli, M., Buitelaar, P., and Sintek, M. (2007). Modeling Linguistic Facets of Multimedia Content for Semantic Annotation. In: *Proceedings of SAMT07 (International Conference on Semantics And digital Media Technologies)*, Genova, Italy, pp 240-251.

Simov, K., Peev, Z., Kouylekov, M., Simov, A., Dimitrov, M., Kiryakov, A. (2001). CLaRK - an XML-based System for Corpora Development. In: *Proceedings of the Corpus Linguistics 2001 Conference.* Lancaster, UK.

Simov, K. and Osenova P. (2007) Applying Ontology-Based Lexicons to the Semantic Annotation of Learning Objects. in: *Proceeding of RANLP 2007 workshop on Natural Language Processing and Knowledge Representation for eLearning Environments.* Borovets, Bulgaria. pp 49–55.

Simov, K. and Osenova P. (2008) Language Resources and Tools for Ontology-Based Semantic Annotation. In: *Proceeding of OntoLex 2008 Workshop at LREC 2008.* Marrakech, Morocco. pp. 9-13.

Vossen P. (1999). Editor. *EuroWordNet General Document. Version 3, Final, July 19, 1999.* http://www.hum.uva.nl/ ewn

Ernesto William De Luca, Andreas Nürnberger

LexiRes RDF/OWL Editor: Maintaining Multilingual Resources

In this article we present the RDF/OWL LexiRes Tool that supports authors using RDF/OWL structures in editing and structuring ontologies. In contrast to standard ontology editors like e.g. Protégé [20], this tool was especially designed to support the use of multilingual resources, and in particular the use of RDF/OWL EuroWordNet word senses and translations [4]. Related concepts can be searched and explored in different languages. Furthermore, concepts can also be merged using automatic or manual merging methods [7]. Thus, the RDF/OWL LexiRes tool supports authors in adding OWL ontologies to the RDF/OWL EuroWordNet representation or to manage the WordNet resources within other external OWL structures.

1 Introduction

Language Engineering involves the development and application of software systems that perform tasks concerning the processing of human natural language [3]. Different tools have been designed, constructed, and are used, for tasks like translation, language teaching, information extraction and indexing. Other, more intangible "language engineering tools" are language resources. Language resources are essential components of language engineering, containing a wide range of linguistic information with different degrees of complexity. These linguistic resources are sets of language data and descriptions in machine readable form, used for building, improving or evaluating natural language and speech systems or algorithms. In, e.g., [2] various types of language resources, i.e. written and spoken language corpora, lexicons and terminological databases are briefly presented.

In the following, we concentrate on lexical resources that provide linguistic information about words. This information can be represented in very diverse data structures, from simple lists to complex repositories with many types of linguistic information and relations attached to each entry, resulting in network-like structures. Lexical resources are used in Natural Language Processing, for example, to obtain descriptions and usage examples of different word senses. Different word senses refer to different concepts, and concepts can be distinguished from each other not only by their definitions or "glosses", but also by their specific relations to other concepts. Such disambiguating relations are intuitively used by humans. However, if we want to automate the process of distinguishing between word senses (word sense disambiguation), we have to use resources that provide appropriate knowledge, i.e. sufficient information about the usage context of a word. One of the most important resources available for this purpose is WordNet [9] and its multilingual variants, including MultiWordNet [22] and EuroWordNet [27].

However, many lexical resources or ontologies, especially WordNet, provide frequently too fine grained word sense distinctions for specific applications like user support in cross-language information retrieval systems. Therefore, we implemented the LexiRes RDF/OWL tool that gives the possibility to navigate lexical information, helping authors of already available lexical resources in deleting or restructuring concepts using automatic merging methods. The restructured information can be navigated and explored. Authors can decide if word senses are unambiguous and important enough to keep them in the hierarchy at the same place or if they express similar concepts and can be merged under the same (now, more general) meaning.

1.1 Outline of the paper

The remainder of this article is structured as follows. The lexical database WordNet and its variants are shortly introduced in Section 1.2. Then, different WordNet representations are discussed in Section 2. Afterwards, we discuss some preliminaries and present our conversion of the EuroWordNet structure in OWL (see Section 3.1). In Section 4, we show the functionalities of the LexiRes RDF/OWL Tool and the extensions already available (see Section 4.2 and 4.3). Some concluding remarks are finally given in Section 5.

1.2 WordNet

WordNet [9, 18] is an electronic lexical database designed by use of psycholinguistic and computational theories of human lexical memory. It provides a list of word senses for each word, organized into synonym sets (synsets), each representing one constitutional lexicalized concept. Every synset is uniquely identified by an identifier (synsetId). It is unambiguous and carrier of exactly one meaning. Furthermore, different relations link these elements of synonym sets to semantically related terms (e.g. hypernyms, hyponyms, etc.). All related terms are also represented as synset entries. These synsets also contain descriptions of nouns, verbs, adjectives, and adverbs. With these information we can describe the usage context of a word.

WordNet distinguishes two types of linguistic relations. The first type is represented by lexical relations (e.g. synonymy, antonymy and polysemy) and the second by semantic relations (e.g. hyponymy and meronymy). Glosses (human descriptions) are often (about 70% of the time) associated to a synset [1]. With WordNet 2.0 also nominalizations - which link verbs and nouns pertaining to the same semantic class, and domain links (based on an "ontology") that should support the disambiguation process have been introduced. Figure 1 represents an example of the ontology hierarchy defined by WordNet [18]. This ontology hierarchy represents different meanings of a given word that are related to it (e.g. meaning 1, meaning 2 and meaning 3) having different linguistic relations (e.g. meaning 1 with hyponym 1.1 and hyperonym 1.1, hyperonym 1.2 and hyperonym 1.3). WordNet can be used for different applications, like word sense identification, information retrieval, and particularly for a variety of content-based tasks, such as semantic query expansion or conceptual indexing in order

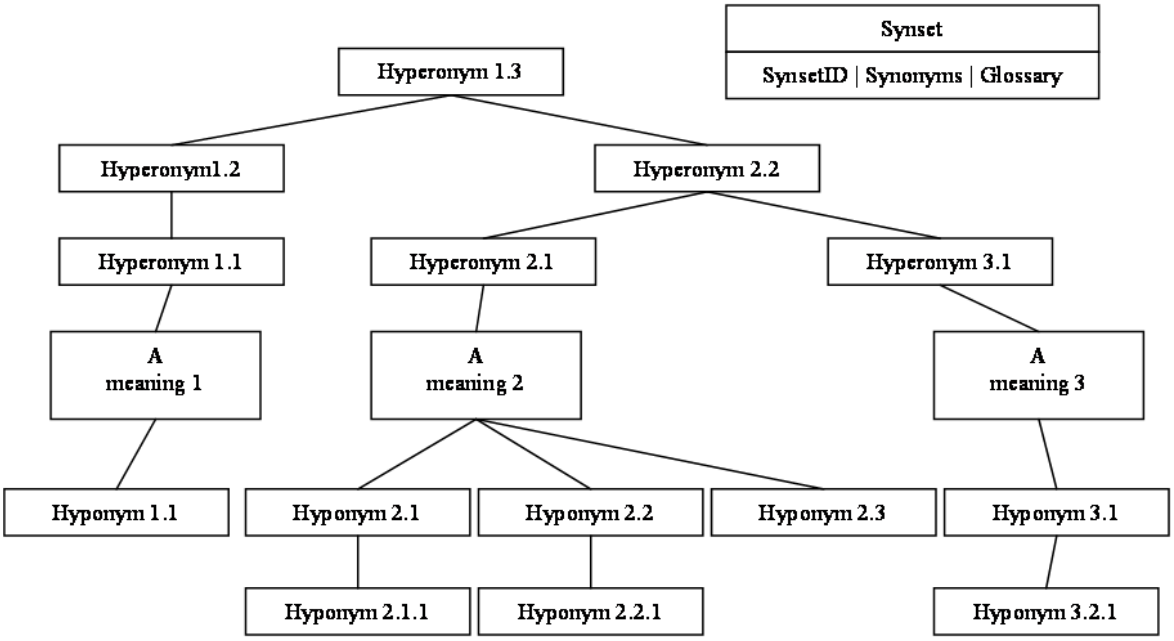

Figure 1: Example of an ontology hierarchy for a given term A.

to improve the retrieval performance [26]. It was first developed only for the English language. Then different versions were developed for several other languages as for example EuroWordNet [27] for several European languages (Dutch, Italian, Spanish, German, French, Czech and Estonian). Its structure is the same as the Princeton WordNet [18] in terms of synsets with different semantic relations between them. Each individual wordnet represents a unique language-internal system of lexicalizations. Furthermore, the Inter-Lingual-Index (ILI) was introduced in order to connect the WordNets of the different languages. Thus, it is possible to access the concepts (synsets) of a word sense in different languages.

1.2.1 Problems of the WordNet Hierarchy

In the following we briefly examine the main semantic limitations of WordNet (and its variants) and describe some problems that have to be solved for its better expressiveness:

- Some lexical links of WordNet should be interpreted using formal semantics in order to express "things in the world". Oltramari et al. [21] revise the Top Level of WordNet (upper or general level) where the criteria of identity and unity are very general, in order to recognize the constraint violations occurring in it. The concepts of identity and unity are described in, e.g., [21].

- The too fine-grained description of synsets is another limitation for its use in natural language processing. Therefore, a restructuring process of the synsets is needed [7].
- Another critical point is given by the confusion between concepts and instances resulting in an "expressivity lack" [10]. For example, if we look for the hyponyms of "mountain" in older versions of WordNet, we will find the "Olympus mount" as a subsumed concept of the word treated as "volcano" and not as instance of it. Thus, we do not have a clear differentiation between abstract descriptions (concepts) and their instantiations (instances). In the newer WordNet Version 3.0 "Olympus mount" has been changed in "Mount Olympus" and explained as instance of "mountain peak", but this problem has not been solved for all other cases. We also have the problem that we cannot use only concepts or only instances because there is no intended separation between them in WordNet.
- Motta et al. [19] treat also the important difference between endurance and perdurance of the entities that should be included in WordNet. Enduring and perduring entities are related to their behaviour in time. Endurants are always entirely present at any time they are present. Perdurants are only partially present, in the sense that some of their proper parts (e.g., their previous phases) may be not present. However, these aspects of instances are not discussed in this article since they seem to be of less importance for the considered disambiguation problem.

When we concentrate on EuroWordNet, we can see that these problems persist, and other problems related to the Inter-Lingual-index and to the language-dependent coverage of word senses come along. Further more detailed discussions of these problems can be found in [10], [11] and [21].

1.2.2 WordNet Related Work

Horák and Smrož [14] developed VisDic for browsing and editing multilingual information taken from EuroWordNet. The tool supports users browsing static information by using only text blocks. Another web interface for multilingual information browsing is presented in [23]. Here a parallel corpus annotated with MultiWordNet [22] can be browsed as well as the words with their related annotated word senses. However, the corpus is very restricted and all accessible information is static. This interface can be used only for a bilingual search in a closed domain. Other work dealing with the lexicography has shown that researchers in this area mostly deal with multilingual lexical resources or corpora only, without the possibility of merging similar word senses.

Given that the EuroWordNet format is defined by the EuroWordNet Database Editor Polaris that uses a proprietary specification, we decided to convert the EuroWordNet Database into a standardised OWL format, in order to access it with standard OWL query tools. In this way we can also enrich this ontology with additional domain-specific ontologies.

ex:car{car:1 auto:1 automobile:1 machine:6 motorcar:1}

Figure 2: Example of the set of nouns of "car" in WordNet 3.0

2 WordNet and EuroWordNet in XML and RDF/OWL

The lexical database WordNet (WN) contains sets of word senses that are synonyms or near-synonyms of each other (synsets). We can also think of a synset as a concept and consider the word senses it contains as (largely interchangeable) linguistic expressions that can be used to refer to it (see Figure 2). Between synsets, semantic relations such as hyponymy (subsumption) and meronymy (part-whole-relation) are defined. For this reason, WordNet has also been called a lexical ontology. Although centered around the synset notion, WordNet also includes additional lexical relations, defined between individual word senses instead of synsets; antonymy (the relation between opposition pairs) is an example. EuroWordNet (EWN) [27] is a multilingual lexical database built along the lines of the WordNet model. In addition to the central relations taken over from WordNet, EuroWordNet offers interlingual as well as further semantic relations. EWN data is distributed on CD-Rom in two formats: plain text and binary files. The binary data can be viewed with proprietary tools.

2.1 XML-based representations

The EuroWordNet data described above has been represented in a proprietary XML format. The XML entries were produced in the BalkaNet project and can be viewed and edited with different tools (VisDic [14] and DEBVisDic [13]). The work on VisDic was motivated by several reasons; among others, distribution of the original tool for viewing and editing EWN data (Polaris) had long been discontinued, and a new platform was needed for the work on BalkaNet. However, a different XML-based format for WordNet has been proposed by the Semantic Web community [25]. Their RDF/OWL representation can be queried and processed by standard Semantic Web tools, thus facilitating the integration of WordNet data into Semantic Web applications.

2.2 OWL

Before describing how we converted EuroWordNet in OWL, we first give a brief introduction of the OWL definition. OWL (the Web Ontology Language) is a formal language used for describing web ontologies. It is written in XML syntax and built using RDF Schemas, basing on a larger vocabulary and stronger syntax than RDF [12]. This enables an exact description of web information and relationships between web information. As a W3C recommendation OWL is seen as standard language for the implementation of the Semantic Web. It is subdivided into three sublanguages: OWL Lite, OWL DL, and OWL Full. OWL Lite is primarily used for creating classification hierarchies and simple constraints. The OWL DL sublanguage makes use of description logics, which form the formal logical foundation of OWL. In order to guarantee

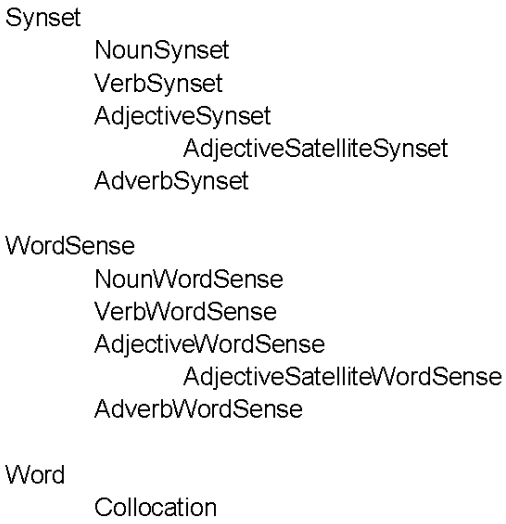

Figure 3: OWL Hierarchy of Princeton WordNet [24].

the representation in first order logic, various restrictions were inserted for the use of RDFs-constructs (for example a class cannot be represented as an instance of another class). OWL Full consists of the same language constructs of OWL DL, but without their restrictions. Hereby ontologies created with OWL can express predicate logic expressions of higher degree.

2.3 RDF/OWL Representation of WordNet

The WordNet Princeton [9] has already been converted into an OWL format as described in [24] using the OWL-DL sublanguage. This representation in RDF/OWL is based on the WordNet data model shown in Figure 3.

When we compare the original Princeton WordNet synset (having only word senses) with the OWL representation, we can see that the RDF/OWL schema (in its full version) has three main classes: Synset, Word and WordSense. The basic version contains only the Synset class. The two classes Synset and WordSense have further four subclasses which are based on the distinction of lexical groups; these are NounSynset, VerbSynset, AdjectiveSynset (which has another subclass AdjectiveSatelliteSynset) and AdverbSynset. The Word class holds the subclass Collocation which denotes terms that are composed of two or more words.

In order to disambiguate the meanings of each instance of a synset, WordSense and Word have a unique URI, that can be used for retrieving words and word senses independently from the synsets. This property was not available in the original version of WordNet. The URIs provide some information about the entity meaning and are built with patterns similar to:

$$\text{wn20instances: } + \text{synset- } + \text{lexical form- } + \text{type- } + \text{sense number.}$$

For example if we want to retrieve the fourth word sense of the word "bank", we would get a URI like: "http://www.w3.org/2006/03/wn/wn20/instances/synset-bank-noun-4."

The properties of the RDF schema are divided into three kinds of relations:

1. those that relate two synsets to each other (e.g. hyponymOf)
2. those that relate two word senses to each other (e.g. antonymOf)
3. and a set of properties that give informations on entities (e.g. XML Schema datatypes like xsd:string as it is used in synsetId).

In order to avoid redundancy, only relations in one transitive direction (e.g. hyponymOf and not hypernymOf) are listed, the others can be retrieved with the owl:inverseOf property implemented in the RDF Schema (see the Figures 6 and 7). Altogether there are 27 relations implemented in the RDF/OWL representation of WordNet. The instances of all classes and properties are separated in several data files, one for the synsets, one for the WordSenses and Words and one for each relation. Although the RDF Schema is used to describe most class and property definitions, there are several OWL statements integrated in the schema to provide better semantical descriptions, like checking the correctness of the data or defining inverse relations. For these statements software have to support the OWL DL standard in order to store and query the data.

3 Representing EuroWordNet in RDF/OWL

Because of the different problems related to WordNet and its variants as briefly discussed in Section 1.2.1, we decided to convert it into an RDF/OWL representation (see below), in order to enable the development of more flexible revision methods. In EuroWordNet, one synset contains all related word senses, synonyms and relations to other synsets and to the Inter-Lingual-Index. This information had to be prepared for inclusion in the appropriate RDF Schema and reorganized for a new data representation. The decision of converting EuroWordNet was also based on the need of extending it (because not all meanings are covered) with other resources. Furthermore, since most domain-specific ontologies are represented in OWL and a WordNet monolingual RDF/OWL representation has already been implemented, a EuroWordNet conversion would add multilingual capabilities to these resources. Therefore, we converted EuroWordNet into an RDF/OWL representation based on the work presented in [24].

Since EuroWordNet has several relations and a structure that are different from the Princeton WordNet, several steps were required to adapt the data to the RDF/OWL Schema of WordNet and to extend this RDF Schema with the new relations. We first analysed the requirements for EuroWordNet and adapted the WordNet RDF Schema to a multilingual representation of EuroWordNet. Then, we converted the EuroWordNet relations into OWL properties and extended the ontology with two domain ontologies [4]. In the following, we describe the steps of this conversion and the problems that

```
<ewn20schema:NounSynset rdf:about="&ewn20instances;synset-bank-noun-1"
    rdfs:label="bank">
    <ewn20schema:synsetId>102690337</ewn20schema:synsetId>
</ewn20schema:NounSynset>
<ewn20schema:Word rdf:about="&ewn20instances;word-bank"
    ewn20schema:lexicalForm="bank"/>
<ewn20schema:NounWordSense rdf:about="&ewn20instances;wordsense-bank-noun-1"
    rdfs:label="bank">
   <ewn20schema:word rdf:resource="&ewn20instances;word-bank"/>
</ewn20schema:NounWordSense>
<rdf:Description rdf:about="&ewn20instances;synset-bank-noun-1">
   <ewn20schema:containsWordSense rdf:resource="&ewn20instances;wordsense-bank-noun-1"/>
   <ewn20schema:containsWordSense rdf:resource="&ewn20instances;wordsense-bank_building-noun-1"/>
</rdf:Description>
```

Figure 4: RDF/OWL-EuroWordnet synset Example.

arose in more detail. An example on how to add ontologies is later given in Section 4.2 based on the OWL pizza and travel ontologies. The RDF/OWL-EuroWordNet Representation has also been extended with linguistic data included in the Hamburg Metaphor Database (HMD) in Section 4.3.

3.1 Conversion of EuroWordNet in RDF/OWL

The steps required to convert EuroWordNet in RDF/OWL can be subdivided into:

- Analysis of the requirements for EuroWordNet
- Adaptation of The WordNet RDF-Schema to EuroWordNet
- Multilinguality
- OWL Property Conversion
- OWL Domain Extension

Van Assem et al. [24] distinguish Word and WordSense in their datamodel for two reasons. First of all, several relations are defined for word senses and synsets and WordNet uses this distinction in its database. Secondly, for the sake of ontological clarity, they assume that synsets include word senses, in order to partition the logical space of the lexicon (words as forms or meanings, and synsets as clusters of word senses by abstracting their distributional context). Agreeing with their model, we adapted their schema to convert EuroWordNet, applying this assumptions also for a multilingual task. An example of an OWL-EuroWordNet synset is given in Figure 4. Here the word sense "bank" is shown within its synset (and synsetId), WordSense, Word and synonyms (containsWordSense).

Analysing the structures of EuroWordNet and of the OWL representation of WordNet, we could recognize that some relations are supported in both versions (see Figure 6). Since EuroWordNet contains relations and properties that are not supported in the WordNet OWL representation, we had to adapt the RDF-OWL Schema to our needs in order to cover these gaps. Therefore, we created and stored new RDF structures, containing these new relations and thus extended the WordNet OWL implementation (see Figure 7). Some other properties that are covered in the WordNet OWL representation (e.g. the property tagCount used in the WordSense OWL declaration) are not available in EuroWordNet, so that we could not consider them.

Because we also tried to avoid redundancy as discussed in [24], we decided to delete the relations in EuroWordNet, which have an inverse form. We compared them updating their inverse relation property, where necessary (see the Figures 6 and 7). This means, if an instance of an inverse relation of a distinct instance is present, the instance can be deleted. If no inverse instance is available the instance is added as inverse relation instance. An example is the hyperonym-hyponym relation resulting in a simple hyponymOf relation. Here the hyponym relation was available in both representations but its name had to be changed from has_hyperonym EuroWordNet format to the hyponymOf OWL-WordNet format (see Figure 5).

Another point to consider was that EuroWordNet contains also different relations belonging to the same "upper relation description" (e.g. ROLE_AGENT, ROLE_INSTRUMENT, ROLE_LOCATION, etc. belonging to ROLE), because of their similar functionality. We decided in this case to merge them all into the same "Upper Relation" RDF file. A similar decision was done in the Princeton Conversion within the pertainsTo relation. The complete mapping between EuroWordNet, OWL-WordNet and OWL-EuroWordNet relations is given in Figure 6. The relations that are only available in EuroWordnet and have been included as new RDF/OWL-EuroWordNet are shown in Figure 7.

Since EuroWordNet is a multilingual resource (and not a monolingual like WordNet), we had to create for every language a unique set of files (containing the language-dependent synsets and relations). Every file of this set was additionally tagged with the name of the corresponding language (e.g. for English: eurowordnet-english-synset.rdf, eurowordnet-english-wordsensesandwords.rdf, eurowordnet-english-hyponymOf.rdf, etc.). We used the available EuroWordNet Inter-Lingual-Index that contains synsets, having the same indentifier (synsetId) for all word meanings in all languages and an illustrative gloss. Because of the redundancy problem already described above, we decided to maintain only the gloss information included in the Inter-Lingual-Index

```
<rdf:Description rdf:about="&ewn20instances;synset-bank-noun-1">
   <ewn20schema:hyponymOf rdf:resource="&ewn20instances;synset-deposit-noun-1"/>
</rdf:Description>
```

Figure 5: RDF/OWL-EuroWordnet hyponymOf Example for the word "bank".

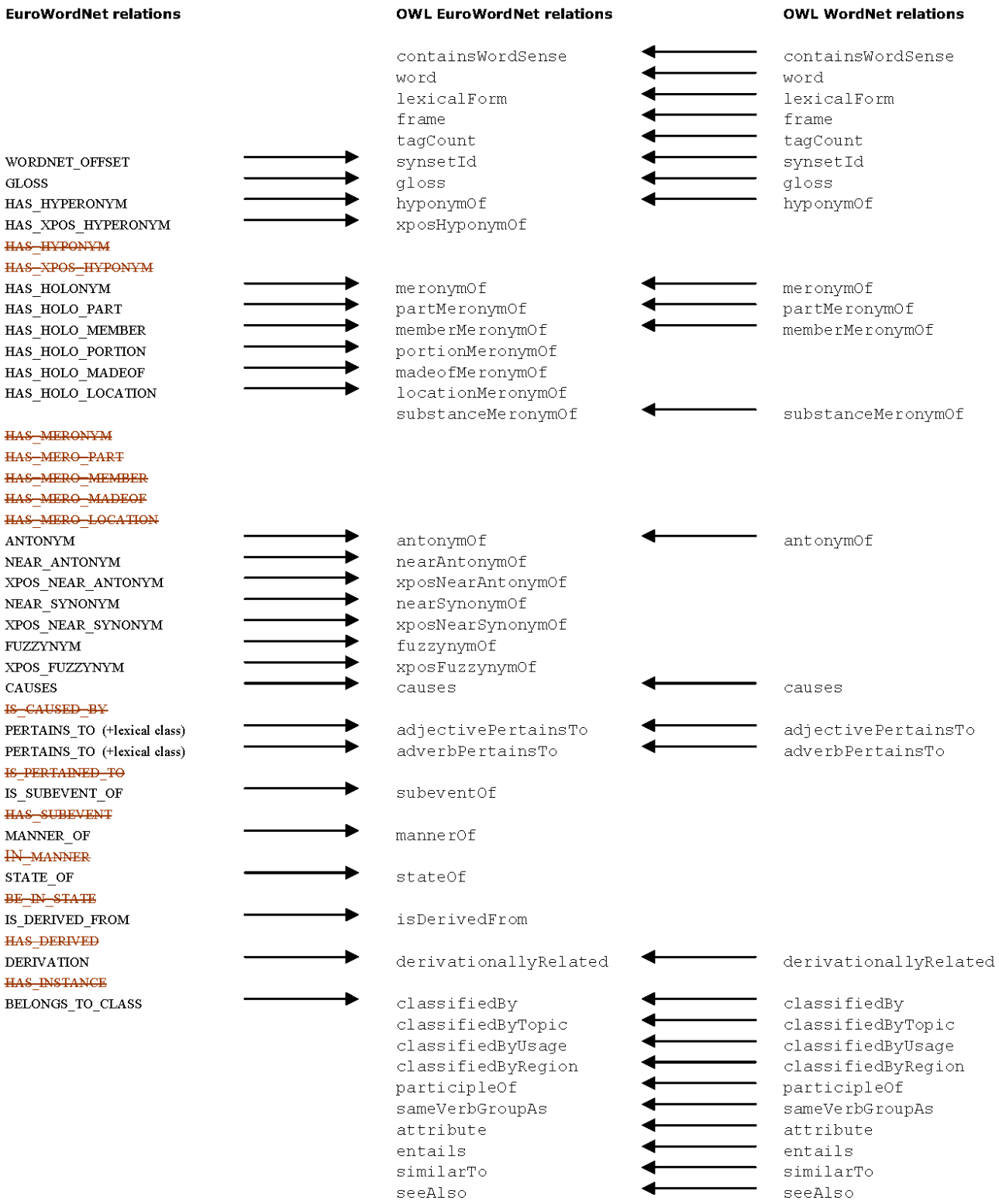

Figure 6: EuroWordnet, RDF/OWL-EuroWordNet and RDF/OWL-WordNet relations. The crossed out relations have been removed, since they are inverse relations to the existing ones.

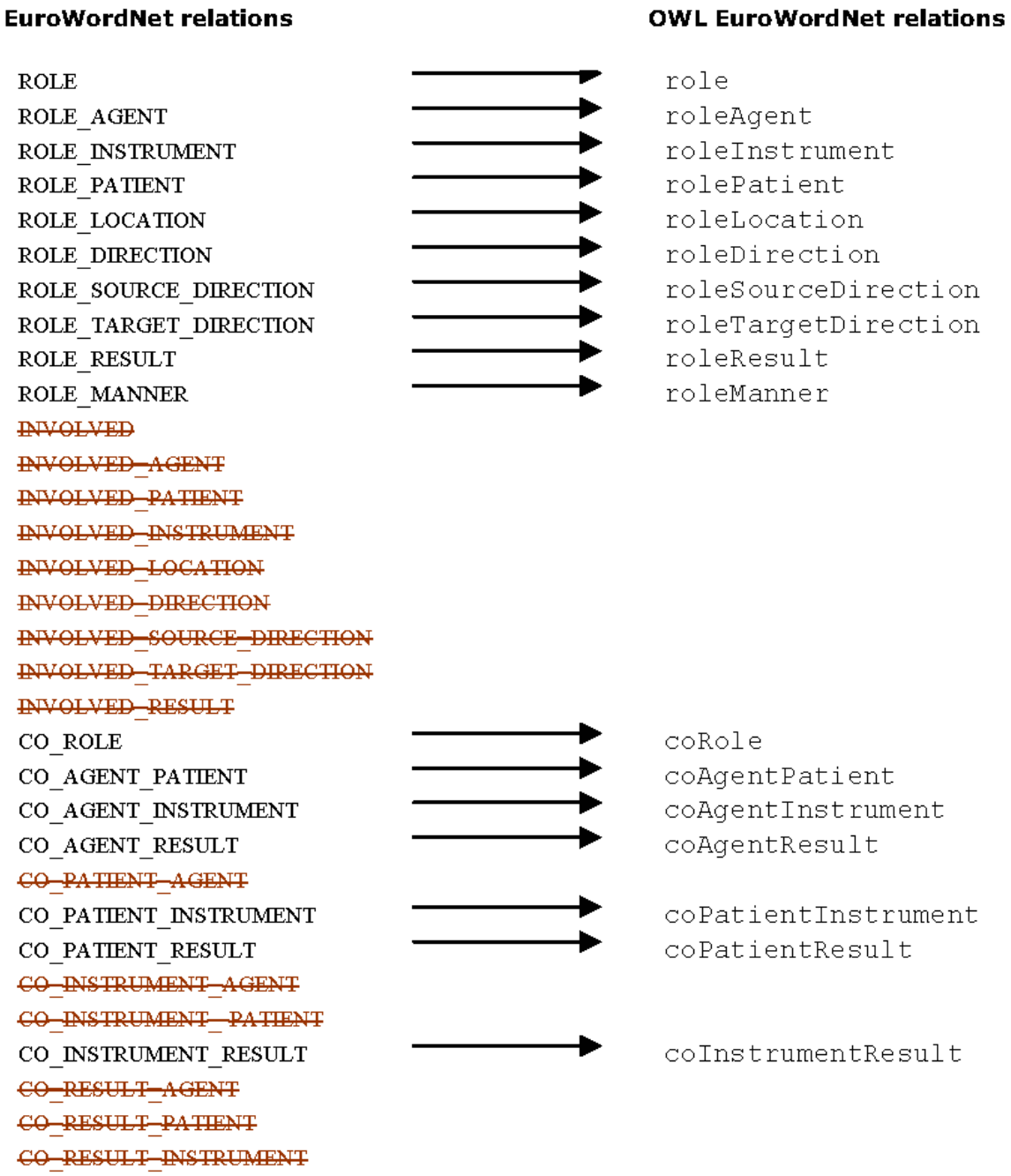

Figure 7: EuroWordnet and RDF/OWL-EuroWordNet relations missing in RDF/OWL-WordNet. The crossed out relations have been removed, since they are inverse relations to the existing ones.

deleting the word senses and synsetIds (already included in the english conversion). Therefore, the gloss entries were extracted and stored in a separate RDF file. Depending on this decision, the synsets of all languages are connected to another through the same identifier (synsetId) describing the same concept in different languages, instead of the Inter-Lingual-Index entries.

4 The LexiRes RDF/OWL Tool

The main idea of the RDF/OWL LexiRes Tool is to give authors the possibility to navigate the ontology hierarchy in order to re-structure it, by manual merging, adding or deleting word senses. The tool is implemented in Java and uses the Jena Semantic Web Framework [17] for querying and retrieving lexical data. It provides an RDF/OWL model in order to access and query the lexical resource. Using EuroWordNet for cross-language retrieval, we support the author in:

- Exploring the lexical resource ontology hierarchy
- Disambiguating the word senses of a query word
- Giving the translations of a query word in different languages
- Creating individual lexical collections
- Adding and deleting meanings
- Merging meanings
- Importing OWL ontologies

Figure 8 shows a screenshot of the LexiRes RDF/OWL editor. On the top left side, we can choose the source language and enter the query term. On the right side (under the "Show Relations" area), we can choose which collection we want to use and which linguistic relations are to be considered for visualization. Query translations can be enabled in the "Show Translations" area. For example, looking for the word "bank", in the English language, the ontology engine retrieves 15 meanings. These meanings describe the different word senses. Every word sense is represented as a synset. The author can choose to "Show Properties" or "Hide Properties" with a left mouse click on a synset. Here all synset-related information is shown. The original RDF resource part of the synset can also be displayed by clicking on the right mouse button and choosing the "Show RDF Resource" option. The properties and the RDF code are then shown on the right-hand side under the "Details" box. After logging in, a user-specific lexical resource collection can be created. In our case, the collection contains a reference to the EuroWordNet lexical resource (as default). The author can add or remove meanings in order to enrich or restructure the hierarchy. It is also possible to query the adapted EuroWordNet lexical resource. To create new meanings, the author has to integrate them into the hierarchy. This is achieved by specifying

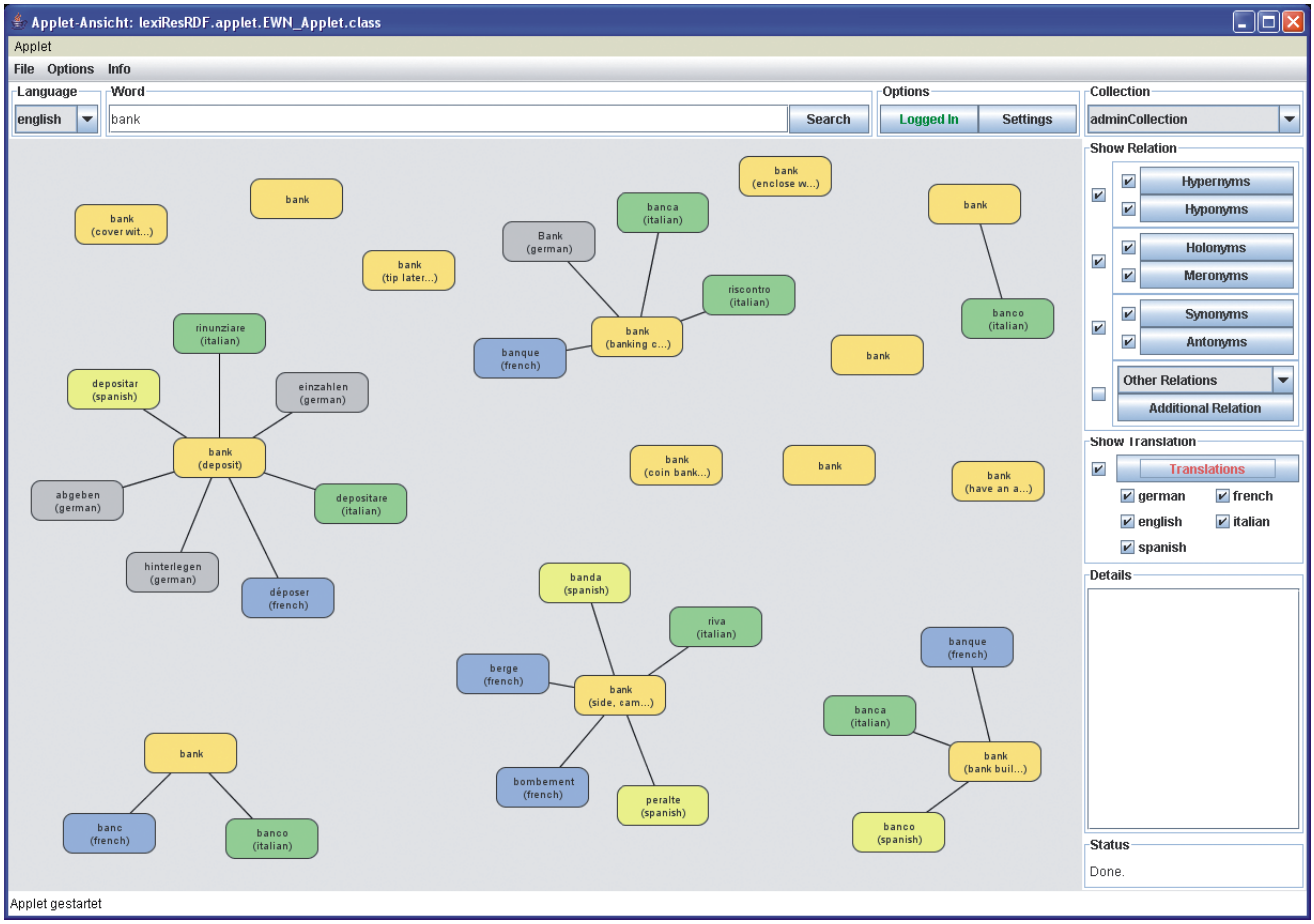

Figure 8: Example of the word “bank” - synset translations - in the LexiRes RDF/OWL Editor.

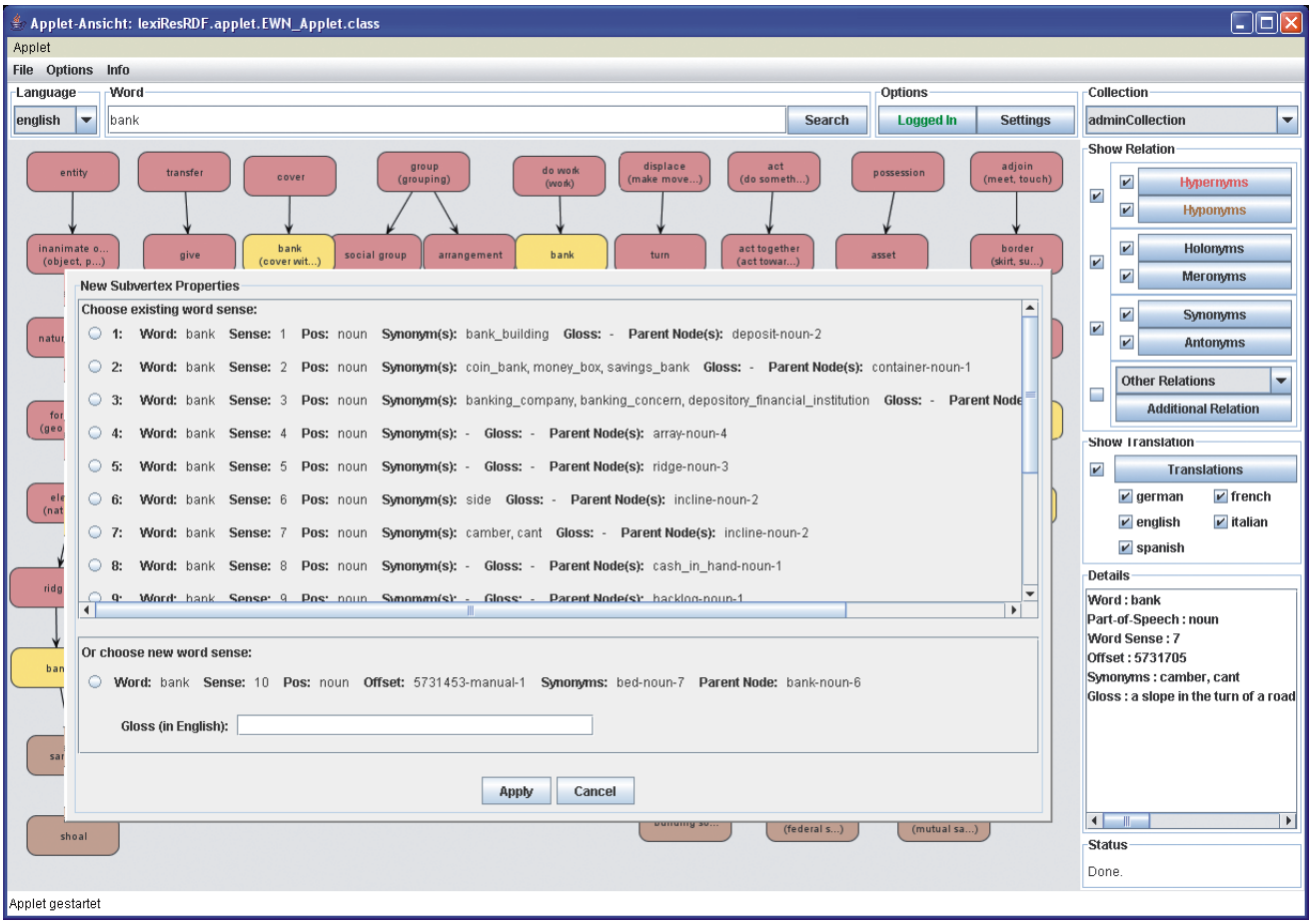

Figure 9: Example of the word “bank” - create new word sense - in the LexiRes RDF/OWL Editor.

the most appropriate superordinate node. New words (and their related terms) can be entered in the "Create New Word Sense" dialog. The system searches for known meanings of these terms and suggests (to the author) a list of candidates with their synonyms, descriptions and generic terms. If any meaning matches the meaning of the query term in the hierarchical context, it can be selected and grouped under the superordinate node. Alternatively, the author can generate a new meaning which is then added to the hierarchy (see Figure 9). External domain-specific ontologies can be merged into the collection using the "Import Ontology" option. Then, the ontology can be uploaded and, if suitable, be added in the relation hierarchy. Further details are given in [4]. When a word sense is removed, the system updates the hierarchy by also removing the respective connections from the linguistic relations. In a graphical representation, this corresponds to deleting all adjacent edges along with the node. If a meaning is deleted, the resulting lack of connection between super- and subordinate words becomes a problem. Because semantic relations do not have to be transitive, the super- and subordinate nodes cannot always be directly connected. Such situations have to be resolved by the author.

The tool also allows the manual merging of synsets when the author decides that two synsets belong to the same meaning and/or describe the same concept. For example, the two "bank" synsets under the superordinate "incline" synset in Figure 10 could be merged. Therefore, the author can pick a "source" synset in the hierarchy that should be merged to a "target" synset. The "Merge Word Sense To ..." menu shows all possible target meanings. The "source" meaning with all its relations is transferred to the "target" meaning.

The synsets can be also automatically retrieved and translated in the different languages available in the ontology (see Figure 8). These can be set within the menu button language and can be shown – always synset-dependent – within a simple click. We can notice that not all synsets have a translation, due to the lexical gaps or the missing entries in the lexical resource.

4.1 Fine and Coarse Grained Representation of Word Senses

The LexiRes RDF/OWL tool gives the possibility to manually merge synsets, when the authors decide that two synsets belong to the same meaning and/or describe the same concept. Authors working with LexiRes can also use an automatically created list of candidate synsets that can be merged. This list can be created with the approaches discussed in [7]. The system proposes the list of changes and the user can select to accept all or check each proposal for merging manually. At the moment these merging methods are implemented outside the tool. The resulting list of possible merging synsets is first checked from the authors and then done manually. After having restructured the ontology hierarchy, a new set of synsets is created. This set is supposed to contain only word senses that are carrier of a distinctive meaning in the context of the considered application. This is a very important step for a use of lexical resources in information retrieval. The possibility to merge synsets in advance

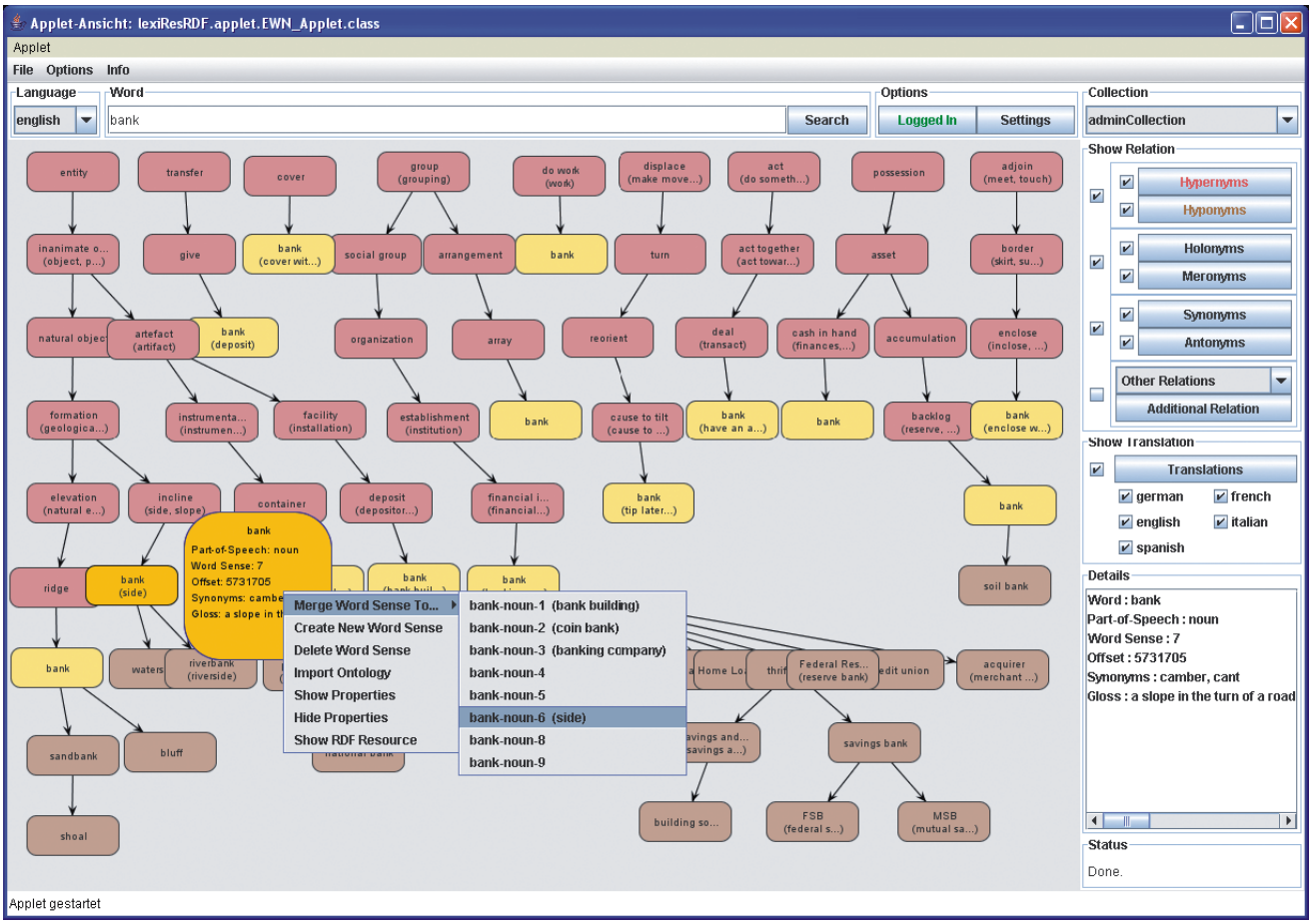

Figure 10: Example of the word “bank” - manual merging functions - in the LexiRes RDF/OWL Editor.

gives the advantage to categorize the retrieved documents disambiguating them with structured word senses that facilitate an automatic classification process [6]. A detailed description of the evaluation of the automatic merging methods applied to the WordNet synsets in given in [8].

4.2 Extending RDF/OWL-EuroWordNet with the *pizza.owl* and *travel.owl* domain-ontologies

After a first conversion of EuroWordNet in an OWL representation, we decided to try to extend it with two OWL ontologies (pizza.owl and travel.owl). For integrating them in our EuroWordnet OWL representation, we analysed their hierarchy of classes that were also built in OWL DL. Every term is declared as a class (owl:Class) and every underlying term as a subclass (rdfs:subClassOf). There are additional restrictions, e.g. owl:disjointWith or owl:someValuesFrom statements. But there are no additional properties (except the OWL DL statements defined). In order to extend EuroWordNet with these ontologies we used a two steps approach:

1. Conversion from the OWL format to the EuroWordNet OWL format

2. Integration of the converted data in the EuroWordNet OWL hierarchy

First of all, we converted all classes (owl:class) into RDF/OWL synset classes (e.g. ewn20Schema:NounSynset), so that they were easier to add into the OWL-EuroWordNet hierarchy. Using some of the merging methods [7], we started a query for finding the synset we wanted to use as hyperonym of the domain ontology to be added (in this example "pizza"). We explored first the hierarchy with the LexiRes RDF/OWL tool. Then, we tried to disambiguate the word senses of the searched word ("pizza") in order to find the correct synset to be extended. After having found the correct synset, we merged manually the complete converted domain ontology under the appropriate hyperonym (synset). In this case we could enlarge the EuroWordNet coverage with domain-specific terms. The extension can be easily recognized (and deleted if needed), because we added to the synsetId the name of the domain ontology followed by the number of word senses (see Figure 11). The pizza.owl ontology, for example, has also a language description for every class (e.g. xml:lang="en"), so that we could add it to the correct set of language files, if available. The same procedure was applied to the travel.owl ontology. A more detailed description of this extension work is given in [4].

4.3 Extending RDF/OWL-EuroWordNet with the Hamburg Metaphor Database

The RDF/OWL-EuroWordNet representation has also been extended with data included in the Hamburg Metaphor Database (HMD), a relational database of French and German corpus attestations containing metaphors [16]. In the HMD, each metaphor is manually analyzed and annotated at several levels: Among other lexical features,

```
<ewn20schema:NounSynset rdf:about="&ewn20instances;synset-vegetarian_pizza-noun-1"
    rdfs:label="vegetarian pizza">
   <ewn20schema:synsetId>5068318-pizza.owl-1</ewn20schema:synsetId>
</ewn20schema:NounSynset>
<ewn20schema:Collocation rdf:about="&ewn20instances;word-vegetarian_pizza"
    ewn20schema:lexicalForm="vegetarian pizza"/>
<ewn20schema:NounWordSense rdf:about="&ewn20instances;wordsense-vegetarian_pizza-noun-1"
    rdfs:label="vegetarian pizza">
   <ewn20schema:word rdf:resource="&ewn20instances;word-vegetarian_pizza"/>
</ewn20schema:NounWordSense>
<rdf:Description rdf:about="&ewn20instances;synset-vegetarian_pizza-noun-1">
   <ewn20schema:containsWordSense rdf:resource="&ewn20instances;wordsense-vegetarian_pizza-noun-1"/>
</rdf:Description>
<rdf:Description rdf:about="&ewn20instances;synset-vegetarian_pizza-noun-1">
   <ewn20schema:hyponymOf rdf:resource="&ewn20instances;synset-pizza-noun-1"/>
</rdf:Description>
```

Figure 11: OWL-EuroWordnet (Merged) “vegetarian pizza” synset Example.

HMD provides references to EuroWordNet synsets. In addition, conceptual information is indicated in terms of domain labels from the Berkeley Master Metaphor List [15].

To provide an RDF/OWL representation of HMD data, we started by defining a new relation between synsets, the conceptual relation extMetaphorOf (“extension by metaphor of ...”). This conceptual relation holds between a synset with a metaphorical meaning and a synset with a literal meaning of at least one of the contained word senses. The relation as such is defined by an RDF schema. We then populated the extMetaphorOf-relation by deriving 107 instances from the HMD data for French. This was done by converting the data concerning attested metaphorical mappings between EWN synsets from the HMD relational database into RDF. The 107 instances of the extMetaphorOf-relation thus represent cases where both the literal and the metaphorical synset were already contained in the original version of EuroWordNet. As with each relation in RDF/OWL EuroWordNet, the resulting information is stored in a separate RDF-file (extMetaphorOf.rdf) and can be distributed as such. A detailed description about the integration of the Hamburg Metaphor Database into the RDF/OWL-EuroWordNet format can be found in [5].

5 Conclusions

In this article we discussed the LexiRes RDF/OWL tool, an editing and visualization tool we developed and adapted for handling multilingual resources (like EuroWordNet) with OWL ontology structures. Furthermore, we described the conversion of EuroWordNet in OWL and presented, based on this, the functionalities of the LexiRes RDF/OWL tool and the already implemented extensions available. This work is a first attempt to evaluate how appropriate EuroWordNet could be represented as OWL on-

tology and how the already existing ontologies could be merged to this representation. The use of this OWL implementation and its performance has to be evaluated in more detail in order to be aware of its benefits and possibly still existing problems.

References

[1] Fabio Ciravegna, Bernardo Magnini, Emanuele Pianta, and Carlo Strapparava. A project for the construction of an italian lexical knowledge base in the framework of wordnet. Technical Report IRST 9406-15, IRST-ITC, 1994.

[2] R. Cole, J. Mariani, H. Uszkoreit, A. Zaenen, and V. Zue. Survey of the State of the Art in Human Language Technology. Center for Spoken Language Understanding CSLU, Carnegie Mellon University, Pittsburgh, PA, 1997.

[3] Hamish Cunningham. A definition and short history of language engineering. Nat. Lang. Eng., 5(1):1–16, 1999.

[4] Ernesto William De Luca, Martin Eul, and Andreas Nürnberger. Converting EuroWordNet in OWL and Extending It with Domain Ontologies. In Proceedings of the Workshop on Lexical-Semantic and Ontological Resources. In Conjunction with the GLDV Conference (GLDV 2007), 2007.

[5] Ernesto William De Luca and Birte Lönneker-Rodman. Integrating Metaphor Information into RDF/OWL EuroWordNet. In European Language Resources Association (ELRA), editor, Proceedings of the Sixth International Language Resources and Evaluation (LREC'08), Marrakech, Morocco, may 2008.

[6] Ernesto William De Luca and Andreas Nürnberger. Improving Ontology-Based Sense Folder Classification of Document Collections with Clustering Methods. In Philippe Joly, Marcin Detyniecki, and Andreas Nürnberger, editors, Proceedings of the 2nd International Workshop on Adaptive Multimedia Retrieval (AMR 2004), part of ECAI 2004, 2004.

[7] Ernesto William De Luca and Andreas Nürnberger. Rebuilding Lexical Resources for Information Retrieval using Sense Folder Detection and Merging Methods. In Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006), Genova, Italy, 2006.

[8] Ernesto William De Luca and Andreas Nürnberger. The Use of Lexical Resources for Sense Folder Disambiguation. In Workshop Lexical Semantic Resources (DGfS-06), Bielefeld, Germany, 2006.

[9] Christiane Fellbaum. WordNet, an electronic lexical database. MIT Press, 1998.

[10] Aldo Gangemi, Nicola Guarino, and Alessandro Oltramari. Conceptual analysis of lexical taxonomies: the case of WordNet top-level. In FOIS '01: Proceedings of the international conference on Formal Ontology in Information Systems, pages 285–296, New York, NY, USA, 2001. ACM Press.

[11] Nicola Guarino and Chris A. Welty. An overview of OntoClean, pages 151–172. Handbook on Ontologies. Springer, 2004.

[12] Ivan Herman, Ralph Swick, and Dan Brickley. Resource description framework (rdf). Technical report, W3C, 2004.

[13] Aleš Horák, Karel Pala, Adam Rambousek, and Martin Povolný. DEBVisDic - first version of new client-server Wordnet browsing and editing tool. In Proceedings of the Third International WordNet Conference (GWC 2006), pages 325–328, Seogwipo, Jeju Island, Korea, January 2006.

[14] Aleš Horák and Pavel Smrož. VisDic - Wordnet Browsing and Editing Tool. In Proceedings of the Second International WordNet Conference (GWC2004), 2004.

[15] George Lakoff, Jane Espenson, and Alan Schwartz. Master metaphor list. Second draft copy. Technical report, Cognitive Linguistics Group, University of California Berkeley, 1991. http://cogsci.berkeley.edu.

[16] Birte Lönneker-Rodman. The hamburg metaphor database project: issues in resource creation. Language Resources and Evaluation, 42(3):293–318, 2008.

[17] Brian McBride, Daniel Boothby, and Chris Dollin. An Introduction to RDF and the Jena RDF API. Technical report, HP, 2006.

[18] George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine Miller. Five papers on WordNet. International Journal of Lexicology, 3(4), 1990.

[19] Enrico Motta, Simon Buckingham Shum, and John Domingue. Ontology-driven document enrichment: Principles and case studies. In KAW'99: 12th Banff Knowledge Acquisition Workshop, Calgary, CA, 1999.

[20] Natalya Fridman Noy, Ray W. Fergerson, and Mark A. Musen. The knowledge model of protégé-2000: Combining interoperability and flexibility. In EKAW '00: Proceedings of the 12th European Workshop on Knowledge Acquisition, Modeling and Management, pages 17–32, London, UK, 2000. Springer-Verlag.

[21] Alessandro Oltramari, Aldo Gangemi, Nicola Guarino, and Claudio Masolo. Restructuring wordnet's top-level: The ontoclean approach. In Proceedings of the Language Resources and Evaluation Conference, LREC2002, 2002.

[22] Emanuele Pianta, Luisa Bentivogli, and Christian Girardi. Multiwordnet: developing an aligned multilingual database. In First International Conference on Global WordNet, Mysore, India, 2002.

[23] Marcello Ranieri, Emanuele Pianta, and Luisa Bentivogli. Browsing Multilingual Information with the MultiSemCor Web Interface. In Proceedings of the LREC 2004 Satellite Workshop on The Amazing Utility of Parallel and Comparable Corpora, pages 38–41, Portugal, 2004.

[24] Mark van Assem, Aldo Gangemi, and Guus Schreiber. Wordnet in RDFS and OWL. Technical report, W3C, 2004.

[25] Mark van Assem, Aldo Gangemi, and Guus Schreiber. RDF/OWL Representation of WordNet. Editor's Draft, 23 April 2006. http://www.w3.org/2001/sw/bestpractices/wnet/wn-conversion.html, 2006.

[26] S. Vintar, P. Buitelaar, and M. Volk. Semantic relations in concept-based cross-language medical information retrieval. In Proceedings of the ECML/PKDD Workshop on Adaptive Text Extraction and Mining, Croatia, 2003.

[27] Piek Vossen. EuroWordNet General Document, Version 3, Final (www.illc.uva.nl/EuroWordNet/docs/GeneralDocPS.zip), 1999.

Sanni Nimb, Lars Trap-Jensen

Use, Re-use and Synergetic Benefit: The Interplay between Wordnet and Dictionary Data

1 Introduction

The first version of a Danish wordnet, DanNet, was released in March 2009. In this paper we will discuss some of the methodological issues of compiling a wordnet on the basis of a large, corpus-based printed dictionary of modern Danish (Den Danske Ordbog, henceforth DDO). Furthermore, we will focus on the synergetic effects between dictionary and wordnet, the first use of DanNet being an onomasiological search engine in the online version of the very same dictionary which was used as the basis for the establishment of the wordnet. The online version of DDO was also published in summer 2009 and can be seen at `www.ordnet.dk/ddo`.

After a short presentation of some basic details about the DanNet project we will describe how the definitions as well as the information on hypernymy in DDO were translated into wordnet relations in DanNet. We will present some typical cases where we decided to encode a semantic feature or relation in DanNet even though, for various reasons, it was underspecified in DDO. Finally, we will show how the wordnet data have contributed to new search possibilities in the online version of DDO.

2 DanNet – background and structure

The Danish wordnet project DanNet is a joint work between two institutions, the Centre for Language Technology (CST) at Copenhagen University, which compiled a pilot version of the computational semantic lexicon SIMPLE-DK for Danish, (Pedersen and Paggio (2004)), and the Society for Danish Language and Literature (DSL), which was responsible for DDO, the dictionary used as the starting point for DanNet (DDO (2005)).[1]

The first version of DanNet contains approximately 40,000 synsets described with hyponymy relations. A large subset of these, approx. 12,000 synsets describing concrete objects or human beings are fully described with a richer set of semantic relations such as meronymy, near-synonymy and antonymy, and in the case of artefacts also relations describing origin, purpose as well as agents and instruments involved in the use of the artefact.

[1] The initial four year phase (2005-2008), which was concluded with the launch of the first version of DanNet, was funded by the Danish Research Council (approx. € 400,000), and an additional funding by the same council of approx. € 130,000 has secured an extension of the wordnet of more than 25,000 synsets as well as links of some 8,000 base concepts to Princeton WordNet, to be ready by the end of 2010.

The wordnet was established on purely monolingual grounds and not, as is the case for many other wordnets, by translating synonym sets from Princeton WordNet to Danish. An important factor in the choice of method – the so-called merge approach – was the fact that a corpus-based dictionary of Danish had been completed in 2005 and was available in a machine-readable version with hypernymy information explicitly specified for each of the approximately 100,000 sense definitions. Thereby, a wordnet for Danish could be semi-automatically built on a well-consolidated sense distinction basis, with sense division and definitions based on corpus evidence, ensuring both a more loyal picture of linguistic conceptualization in Danish and a better sense coverage when it comes to the future computational treatment of Danish text material by the use of DanNet.

3 Use: Danish monolingual dictionary → wordnet for Danish

Asmussen et al. (2007) gives a detailed description of the microstructure of DDO and of the information extraction from the dictionary to DanNet. The article focuses on the semi-automatic generation of the wordnet's hyponymy structure and on the exploration of automatic procedures for assigning other semantic relations, such as meronymy and holonymy, mainly based on semantic patterns extracted from the definitions.

Centrally involved in the generation of a hyponymy structure built on dictionary data was the general and sometimes challenging task of harmonising the raw, automatically extracted hyponymy structure that was derived directly from the genus proximum information contained in DDO. Among the problematic cases was the arbitrary choice of genus proximum in DDO in case of synonymous alternatives. The task of disambiguating cases of homonymy and polysemy was described in Asmussen et al. (2007), but the conclusion was that the compilation of the hyponym hierarchy was indeed facilitated by the utilization of genus proximum information available in the DDO. Finally, it was argued that in order to ensure the practical use of wordnets as resources in formal ontologies, one should be able to separate the so-called taxonomical hyponyms, e.g. for the concept 'tree' the different kinds of trees such as *birch*, *oak* and *cherry tree*, from those hyponyms that are not taxonomical kinds but instead denote, say, a functional aspect as in the case of *roadside tree* or *climbing tree*. This means that an *orthogonal*, i.e. non-taxonomical, feature is added to the hyponyms in DanNet. This information is not directly deducible from the data in DDO which does not, as it is customary in a semasiological dictionary, include information about different groups of hyponyms. Apart from having a more precise set of hyponymy relations than DDO, DanNet also contains information about hyponym categories. See also Pedersen and Sørensen (2006), Pedersen and Nimb (2008) and Pedersen et al. (2009) for further discussion on the subject.

With respect to the automatic extraction of semantic relations, Asmussen et al. (2007) concluded that due to the lack of obligatory structural templates for definitions in DDO regulating both the defining vocabulary and the grammatical expressions, it was not possible to automatically extract semantic patterns on the basis of definitions. In order

to lend itself to automatic extraction, definitions would have to be constructed in a more consistent and predictable way than it was done in DDO, with an explicitly defined syntax connecting certain syntactic patterns with semantic relations, it was claimed. Furthermore, it added to the problem of automatic extraction that no specifications had been prepared to determine which types of features or relations should obligatorily apply to which types of words in the dictionary.

We will discuss how these factors may also cause problems when the DDO definitions are manually translated into wordnet relations, as well as other problems deriving from the use of a dictionary as a lexical knowledge basis for a wordnet.

4 Semantic Relations in DanNet

As described in Asmussen et al. (2007), the definitions of some groups of words in DDO, e.g. in the cases of meals, cakes etc., cover the semantics that we estimate to be relevant in a wordnet and the translation from dictionary definition to semantic relations in DanNet is fairly straightforward in these cases. But as the wordnet compilation proceeded, it became clear that DDO sometimes falls short of the semantic requirements needed in a wordnet to be used for information retrieval, text understanding etc. As a starting point, we will present an overview of the set of semantic relations used in DanNet.

The set of semantic relations in DanNet are identical with the standard WordNet relations, with a few but, in our view, important extensions taken from the Danish SIMPLE project. As opposed to other wordnet models, the SIMPLE model (Lenci et al. (2000)), organized the semantic relations according to the four so-called qualia roles (Pustejovsky (1995)), which relate to inheritance structure, origin, composition and purpose, respectively. None of the standard WordNet relations cover the origin dimension, and the same is true of the purpose dimension of a concept. Our experience from the work on the Danish SIMPLE lexicon (Pedersen and Paggio (2004)) was that all four dimensions of the qualia structure were needed to describe a concept and in fact it was the only effective way to ensure coverage of a maximum number of word sense nuances in the encoding process. Therefore, it was decided to organise the set of relations in the same way in DanNet. So in DanNet, the SIMPLE relations MADE_BY and USED_FOR (telic relations) and the more flexible relation CONCERNS (constitutive relation) were added to the sets traditionally used by wordnets. Finally, we added a USED_FOR_OBJECT relation, used to describe the object of the USED_FOR relation. An example is *brødkniv* 'bread knife' which has the USED_FOR relation *skære* 'cut' and the USED_FOR_OBJECT relation *brød* 'bread'. See Table 1.

The four-sided organisation of the relations was very useful in the establishment of templates for each ontological type in DanNet. E.g. the template type [ARTIFACT+OBJECT] would contain the MADE_BY relation as well as the USED_FOR relation since these will always be relevant in the semantic description of artefacts: they are always man-made and they are always made to be used for a purpose. The interface editor used in the encoding of DanNet produces a synset on the basis of templates. Once the editor has

Formal Role (INHERITANCE)	**Agentive Role (ORIGIN)**	**Constitutive Role (COMPOSITION)**	**Telic Role (PURPOSE)**
HAS_HYPERNYM	MADE_BY (SIMPLE)	HAS_HOLO_MADE_OF	USED_FOR (SIMPLE)
HAS_HYPONYMS		HAS_HOLO_PART	USED_FOR_OBJECT (DANNET)
IS_A_WAY_OF		HAS_HOLO_MEMBER	ROLE_AGENT
		HAS_HOLO_LOCATION	ROLE_PATIENT
		HAS_MERO_MADE_OF	
		HAS_MERO_PART	
		HAS_MERO_MEMBER	
		HAS_MERO_LOCATION	
		CONCERNS (SIMPLE)	
		INVOLVED_AGENT	
		INVOLVED_PATIENT	
		INVOLVED_INSTRUMENT	

Tabelle 1: Semantic relations in DanNet

decided on the ontological type of the concept to be encoded, the tool establishes a synset containing the relations that are relevant for the ontological type in question. In this way, the initial creation of each template encompasses the specifications for those relations that are obligatory for a certain ontological type of words in DanNet. This is the type of specifications that were missing in the dictionary-making process of DDO and one of the main reasons why an automatic extraction of relations was difficult to carry out (Asmussen et al. (2007)).

Furthermore, the encoding process in DanNet begins with the linguistic top hypernyms, e.g. concepts like *bog* ('book'), *legetøj* ('toy') or *beklædningsgenstand* ('garment') with a maximum of semantic relations (see Table 2). In that way, the top hypernym also comes to function as a kind of specification for all types of books, toys and garment, the inheritance mechanism of the DanNet interface ensuring that the full set of relations inherited from the top is visible for all hyponym synsets which are to be described. The job of the lexicographer is to adjust the relations according to the hyponym in question, based on the definition in DDO. E.g. for the concept *kogebog* ('cookery book'), in DDO defined as (in translation): "book containing recipes and sometimes instructions for cooking", the relation inherited from *bog* ('book') CONCERNS: *emne* ('subject') is changed to CONCERNS: *mad* ('food'). Furthermore, a CONCERNS relation is added to the inherited one: CONCERNS: *madlavning* ('cooking'), as well as the relation HAS_MERO_PART: *madopskrift* ('recipe').

In DanNet, a systematic top-down working process is employed, beginning with the two sets of spefications for obligatory relations for various types of words: 1. the templates defined for each ontological type, and 2. the complete sets of relations defined at the top hypernym level and subsequently inherited by all the hyponyms.

	bog (book)	**legetøj (toy)**	**beklædningsgenstand (item of clothing)**
Ontological type	LANGUAGEREPRESENTATION + ARTIFACT + OBJECT	ARTIFACT + OBJECT	GARMENT + ARTIFACT + OBJECT
Formal role INHERITANCE	HAS_HYPERNYM: genstand ('object')	HAS_HYPERNYM: genstand ('object')	HAS_HYPERNYM: genstand ('object')
Agentive role ORIGIN	MADE_BY: skrive ('write'), trykke ('print')	MADE_BY: fremstille ('produce')	MADE_BY: sy ('to sew')
Constitutive role COMPOSITION	HAS_MERO_MADE_OF: papir ('paper') HAS_MERO_PART: tekst ('text'), side ('page'), ryg ('back'), titel ('title') CONCERNS: emne ('subject') INVOLVED_AGENT: forfatter ('writer'), læser ('reader')	INVOLVED_AGENT: barn ('child') HAS_MERO_MADE_OF: materiale ('material')	HAS_HOLO_PART: påklædning ('dressing', 'clothes') HAS_HOLO_LOCATION: kropsdel ('bodypart') HAS_MERO_MADE_OF: stof ('fabric') INVOLVED_AGENT: person ('person')
Telic role PURPOSE	USED_FOR: læse ('to read')	USED_FOR: lege ('to play') USED_FOR_OBJECT: leg ('game')	USED_FOR: klæde ('to dress') USED_FOR_OBJECT: person ('person')

Tabelle 2: Relations on top hypernyms *bog* ('book'), *legetøj* ('toy') and *beklædningsgenstand* ('item of clothing')

This approach is opposite to the bottom-up process adopted in DDO where the main purpose was to present well-formed definitions with a fairly simple syntax intended for a human reader. In DanNet, on the other hand, the aim is to describe, explicitly and as accurately as possible, the semantics of a linguistic item (a sense) in terms of relations to other concepts, taking into consideration the computer programs' complete lack of knowledge of how to use the dictionary. Veale and Hao (2008) claim that not even the kind of knowledge embodied in dictionaries covers what is needed to make a computer understand everyday language. It is argued that wordnets should be enriched with information on stereotypes and culturally inherited associations, e.g. that snakes are related to treachery and slipperiness and that elephants have a good memory, in order to make this possible. This clearly lies outside the scope of DanNet at its current stage. Our aim in DanNet has been to reach an information level defined as 'the native speaker's lexical knowledge about a concept', focusing on the prototypical semantic aspects. In that respect, the ambition regarding information level does not differ from the goal of DDO. The real difference between the two types of lexical resources lies in the fact that a dictionary definition leans heavily on the reader's ability to make assumptions without any explicit statements in the text (Svensén (1993)). Human readers constantly make assumptions and use their knowledge of the world, and compilers of dictionaries base their definitions on this, whether they are aware of it or not, quite contrary to the case of wordnet compilers who must be careful to avoid inferences of any kind and describe everything explicitly in their encodings. Another difference is the syntactic limits of the type of dictionary definition chosen in DDO: most definitions consist of one well-formed, not too complex phrase aiming to capture the typical usage. This style was deliberately preferred over the definitional style of its predecessor, the Danish monolingual dictionary *Ordbog over det danske Sprog* 1918-1956 (ODS (1956)), whose definitions are often quite complex, with frequent use of subordination, parenthetical elements and interposed reservations, alternatives etc. Thirdly, one should keep in mind that a dictionary's sense description is not necessarily confined to the definition only, but may be distributed over several elements. This means that the user will have to read collocations, valency patterns, citations and other relevant data in order to grasp the semantic description in its entirety.

It is therefore hardly surprising that we sometimes find a discrepancy between the actual number of relations described in the definition of a word, and the number of relations which from a systematic point of view should be described for a given word in DDO in order to reflect the native speaker's lexical knowledge.

Bearing this in mind, we will now turn to discussing some typical instances where DDO does not contain sufficient information to specify all the semantic relations required by a certain type of synset in DanNet.

5 'Missing' lexical information in DDO: reasons and general tendencies

In the case of *bog* ('book'), Table 2 lists a series of semantic relations which are considered relevant for the description of the synset in DanNet. Now, compare the list with (the

translation of) the definition in DDO: "printed or written sheets of paper bound or fastened together so as to form a whole, often a coherent text intended for reading". Although probably acceptable to most readers, it is nevertheless striking that the definition says nothing about the writer, nor about the reader, the title and subject of the book. There is no mention of a back, and instead of the common word "page", it has "sheets of paper". These semantic features have all been added in DanNet as they are considered central semantic aspects in the description of *bog* ('book') and its many hyponyms. In fact, the process often turns out to be recursive as relations are added in a series of repeated steps until the final number is fixed at the top level hypernym. The reason is, of course, that it may not be discovered that a relation is relevant until a hyponym is described at a lower, more specific level. A case in point is the CONCERNS relation of *bog* ('book'). Not until the many hyponyms of *bog* were considered, did it become clear that the subject is a central semantic aspect of the concept even though it might not, initially, seem crucial for the concept 'bog' itself: *Koranen* ('the Koran') CONCERNS: islam; *fuglebog* ('bird book') CONCERNS: bird, *kogebog* ('cookery book') CONCERNS: cooking, *kriminalroman* ('crime novel') CONCERNS: crime, etc.

The definitional style of using single well-formed sentences is obviously the main reason why DDO has not been able to include all semantic aspects of the quite complex concept of 'book'. The wordnet model includes much more detailed information than found in DDO and allows many types of relations that may even be used more than once if needed. The same is also true of less complex definitions. For example, for some hyponyms of *bog* we find DDO definitions where central semantic aspects are neglected, probably due to the lack of specifications for the type of words in question: in the definition of *salmebog* ('hymn book'), "book which contains a selection of hymns", nothing is said about the typical user (the church goer), or the typical use (to be sung during a church service). Yet much of the information appears in the entry *salme* ('hymn'). Or consider the example *letlæsningsbog* ('easy reader') where the DDO definition ("book with a typography and a language style that are adapted to the user's low level of reading proficiency") makes no mention of the typical user: a pupil. In this case the information does appear, but only indirectly, in the citation. Another example is the word *butik* 'shop', defined as "room or building where a tradesman displays and sells products". Next to the definition, and without explanation, are a number of collocations, among others *se på butikker* (literally 'to look at shops', i.e. 'to do window shopping') and a citation which translates: "We walked down the pedestrian street Strøget. Mona stopped in front of almost every shop. She loved looking at clothes". The example shows how the dictionary relies on knowledge which is not made explicit in the editorial text. Neither the collocation nor the citation would make sense were it not for the fact that human users know that shops have windows with goods on display. This fact is simply implied in DDO as the definition of 'butik' says nothing about it. A similar case is *indlæggelsesseddel* ('referral note') defined as "document issued by a doctor prescribing hospitalization", and supplied with the citation "The doctor gave Marie a referral note to the sanatorium". Here we find nothing in the definition about the person involved,

i.e. the patient. The human dictionary user has no problem in deducing that Marie is a patient, but in DanNet we must add this explicitly, which is why the concept has been encoded with the relations: HAS_HYPERNYM: document, MADE_BY: to issue, INVOLVED_AGENT: doctor, INVOLVED_PATIENT: patient, USED_FOR: hospitalization.

The process of adding full sets of semantic relations to approximately 6,8oo object artefacts in DanNet has revealed that the telic role, the USED_FOR relation, is usually decribed in DDO. This has confirmed us in our decision to add the USED_FOR relation from SIMPLE to the standard WordNet set of relations. This role is centrally involved in the description of the semantics of artefacts as already pointed out in the SIMPLE project. Another general tendency, especially in the case of complex concepts, is that the definitions often lack information about the parts of the object (book: back, page, shop: display window), even when there is a close lexical, or indeed morphological, link between the parts and the whole. On the other hand, we suspect that the definitions of the parts contain information about the whole more often than vice versa, but at this point it remains a hypothesis as this group of words has not yet been supplied with the full set of semantic relations in DanNet. The main reason for not incorporating all parts of an object in DDO is primarily the demand for well-formed dictionary definitions. Especially in the case of complex objects, too many phrases are needed to provide a comprehensive description.

Finally, we have found a strong tendency not to mention the typical user of an artefact object in DDO (e.g. easy reader: pupil, hymn book: church goer), even in cases of morphological relationship. This could be taken to suggest that the typical user – or producer – of an artefact is not central to the understanding of an object. Instead, the user may be more closely connected to the verb describing the act of use. However that may be, it is interesting that the artefact is often morphologically closely related to the user: shop/shopkeeper/shopper, pharmacy/pharmacist, bakery/baker, pilot licence/pilot. Table 3 shows some examples of typical users, added as relations in DanNet. In all cases, this information is missing in DDO.

To sum up, compared to the information in DDO, DanNet has been extended with highly structured data on hyponymy relations as well as on the type of hyponymy relation (taxonomical or non-taxonomical). Furthermore, a large number of semantic relations that are not mentioned in the definitions of DDO have been added in the case of artefacts, especially information about the parts of the artefact and about the typical user. In a long perspective, the enriched wordnet may in turn be utilized to improve search facilities in an online version of DDO. The next chapter will describe how the first version of the online DDO makes use of the DanNet data to present onomasiological information.

6 DanNet data and the dictionary

It is not a new idea to use wordnet data for human users to present onomasiological information. Various visual representations of the Princeton WordNet are available on the net, such as `www.visualthesaurus.com`, `www.thefreedictionary.com` and `thesaurus.`

Synset	Added in DanNet as compared to DDO
bog ('book')	INVOLVED_AGENT: forfatter ('writer') INVOLVED_AGENT: læser ('reader')
flyvecertificat ('pilot license')	INVOLVED_AGENT: pilot ('pilot')
briller ('glasses')	INVOLVED_AGENT: person ('person')
forskningsbibliotek ('research library')	INVOLVED_AGENT: forsker ('researcher')
læbestift ('lipstick')	INVOLVED_AGENT: kvinde ('woman')
barberkost ('shaving brush')	INVOLVED_AGENT: mand ('man')
ægteskab ('marriage')	INVOLVED_AGENT: ægtepar ('married couple')
apotek ('pharmacy')	INVOLVED_AGENT: apoteker ('pharmacist')
bageri ('bakery')	INVOLVED_AGENT: bager ('baker')
registreringsattest ('vehicle registration certificate')	INVOLVED_AGENT: motorkontor ('motoring office')

Tabelle 3: Relations added in DanNet

`reference.com`. Some of these also offer thesaurus information in combination with dictionary data from one or more dictionaries. A case in point is TheFreeDictionary which has thesaurus information from two sources, Princeton WordNet 3.0 and Collins Essential Thesaurus, as well as dictionary information extracted from The American Heritage Dictionary of the English Language, Collins Essential English Dictionary, bilingual learner's dictionaries from Kernermann Publishing, and various lists of technical terms. The solution is a portal-like presentation where one query is performed in different lexical resources simultaneously and the result is shown as a sequence of adjacent matches.

In the wordnet world, DanNet is unique in that the encoded relations are so closely connected with the dictionary data of DDO. As we have seen, most of the encoding task of the DanNet editors consists of extracting information from the dictionary articles and making the relevant relations explicit, whether they are already expressed directly in the articles, or can be deduced by the human user. In this perspective, the two can be viewed as one combined lexical resource from which both dictionary and wordnet data can be drawn and shown in a user interface. Rather than showing the results of simultaneous queries of two databases, it is our aim to provide one integrated access that offers a choice between a semasiological and an onomasiological presentation of the same underlying data. Let's turn to see how this works in practice.

Figure 1 and 2 show extracts from a DDO dictionary article from DDO as it looks in a prototype version of the user interface. The use of DanNet information becomes

Abbildung 1: Section of the interface with semasiological and onomasiological search option

Abbildung 2: Section of the interface showing ‘related words’

apparent at various places in the article. First, as a quite general possibility, the user has the option of selecting either a semasiological or an onomasiological search mode, displayed in the upper left corner. The traditional (semasiological) view of the dictionary article is selected as default.

Secondly, the dictionary article has been supplied with a new element. In addition to the word relations SYNONYM, ANTONYM and CONFER, already present in the (printed) dictionary, there is a new element RELATED WORDS FOR. Here we have automatically extracted and displayed the hypernym, hyponyms, and co-hyponyms, i.e. synsets sharing the same hypernym as the sense described (in this case the hypernym for *damage* is 'deterioration'), on the assumption that sister senses, although clearly not synonyms, are nevertheless relevant for a user seeking help to produce a text as they often have the same paradigmatic properties as the entry word in a given sense. The information is more helpful for native speakers than for learners of Danish as no explanation is provided to distinguish the words on the list. For practical reasons, only a limited number of co-hyponyms are shown, but the full list unfolds when it is clicked.

Finally, the user can move freely between the two presentational modes. If the user clicks on the icon to the right of the definition (the button with the letter 'B' inside – *b* alluding to *begreb* 'concept'), the interface changes to the onomasiological view of the sense. Clicking on the icon is equivalent to performing an onomasiological search for that word and subsequently selecting the relevant sense from the search result presented in the right column. The outcome of an onomasiological search for *bil* 'car' can be seen in Figure 3. Notice that the search result is the entire synset, not just the synset member *bil* alone. The search term is highlighted, however, and after a definition of the synset (taken from the dictionary) the six members of the synset are listed, each member being clickable if the user wants to view the corresponding dictionary entries.

The main focus of attention is on the hyponym hierarchy, partly for the practical reason that this part of DanNet is the most thoroughly developed, but more importantly because we believe it to be very useful to the human user, given that the most relevant other word relations (synonyms, near-synonyms, antonyms) may be imported directly into the dictionary article and shown in the traditional view. The idea behind the visual presentation is to show the hierarchy in relation to the chosen synset. This synset is displayed as the basic level centrally on the screen and at this point only the levels immediately next to the basic level are visible, in upward and downward direction respectively. The interface solution represents one answer to the conflict between two incompatible ambitions: an intention to show all the details of the hyponym hierarchy, and the wish to help the user keep a sense of orientation and overview.

The same kind of compromise between detail and overview is also the cause of some of the other features that are used. For example, some synsets have a large number of co-hyponyms, and for that reason only a limited number is shown, but the user can choose to see them all with a click. Moreover, as default, the subordinate level is not directly visible because one or more of the co-hyponyms shown can have substantial numbers of hyponyms, thereby causing the users to lose track of where they are in the hierarchy. Instead, they are given the option to click on one of the arrow buttons to

Abbildung 3: Result of an onomasiological search for *bil* ‘car’

unfold the subordinate level for that particular synset. The length of the arrows visually reflects the number of hyponyms: one dash indicates a small number, and two dashes a large number.

Even though only the next immediate level is visible, the user can move up and down in the hierarchy by clicking on a synset. Whenever a synset is selected, the interface updates to display that synset at the basic level and its corresponding immediate levels. To illustrate, we can look at *bil* 'car' again. If the user clicks on the hypernym *motorkøretøj* 'motor vehicle', the interface displays the synset *motorkøretøj* at the centre as the new basic level and shows the hypernym *køretøj* 'vehicle' at the level immediately above. In this way, users can navigate all the way to the topmost level and, similarly, to the most specific item in downward direction. To facilitate navigation and help the users maintain an overview, we have decided to show the entire hierarchy of ancestors in an independent section at the bottom of the central field. This gives the users a quick overview and allows them to jump to non-adjacent levels in the hierarchy should they wish to do so. Because of the frequent branchings at subordinate levels, it should be obvious that a corresponding overview of the hierarchy from the basic level downwards is not feasible.

Another feature, which we have introduced to help with the overview, can be seen immediately before the chosen synset in the hierarchy. Here the user has the option of grouping hyponyms together according to different criteria. This is particularly helpful when the number of hyponyms is large, as is the case for synsets such as *person*, *part* or *place* and *make*, *be* or *get*, which include hundreds or even thousands of synsets as their hyponyms. We have found it necessary to provide the users with some kind of meaningful subgrouping if they are not to become completely lost in the sheer quantity of details. At the same time, the grouping criteria must be of such a nature that they can be employed automatically. The final test deciding which parameters will be implemented has not been concluded at the moment of writing, but the most promising ones are a subset of the DanNet encodings: PURPOSE (the relation USED_FOR), PARTS (the relations HAS_MERO_PART and HAS_HOLO_PART), MANUFACTURE (the relation MADE_BY) and ontological type (ONTOTYPE). The assumption is that synsets which have the same encoding for one of the parameters are likely to have something conceptual in common, and therefore it is meaningful to group them together. However, it cannot be predicted universally which parameter is most meaningful; it varies from synset to synset and must be chosen for each individual lexical unit. Therefore, the grouping feature makes certain demands on the user's ability to make reasonable judgments as they must select the most appropriate parameter themselves in order to profit from this feature.

Although we find the hyponom hierarchy more interesting for human users, we see no reason to deny the user access to the other types of encoding. Relations other than hypernyms/hyponyms can be seen under the heading *Betydningstræk* 'features of meaning' as shown in Figure 4.

It appears that the synset *bil* 'car' has the ontological type [MEANS OF TRANSPORTATION + ARTIFACT], and two relations have been encoded in addition to the hypernym.

Abbildung 4: Features of meaning

The shades of grey of the bars in Figure 4 correspond to different colours on the screen. The first is the holonymic relation PARTS where two parts have been encoded: a car has wheels and it has an engine. The USED_FOR relation also has a dual encoding: it is used for a) transportation or b) driving. The transportation purpose is further subdivided into i) goods and ii) persons or individuals.

In our opinion, information about the other relations is primarily requested by language specialists, and for that reason the panel is hidden in the default view but can easily be unfolded with a click. Notice that the relations can also be used as hints concerning the grouping of hyponyms. In the case of motor vehicles, it seems sensible to group synsets according to their part-whole relation: vehicles that have engines, vehicles with wheels etc. Likewise, the ontological type is often a sensible grouping criterion: The use of ontotype as a grouping parameter will separate, for example, plants that are edible from those that are not for all hyponyms of 'plant, vegetable'. However, it may not always be easy for the user to realise it, but as a first step in that direction it helps to present the ontological type [PLANT+OBJECT+COMESTIBLE].

7 Future perspectives

In the first version of the interface, we have given priority to the details of the DanNet encodings and the user's possibility to move about in the hierarchy. Like other websites, however, we are also contemplating a visual presentation of the same data where several relations can be incorporated at a stroke, thereby facilitating the user's overview of a word's relations to other lexical units, as illustrated in Figure 5.

This idea is not new. It has been implemented in other interfaces using information from Princeton WordNet, for example in The Visual Thesaurus (`www.visualthesaurus.com`) which offers a presentation along similar lines. An example illustrating the same word is given in Figure 6.

Notice that The Visual Thesaurus draws on not only synset based information such as hyponyms, meronyms, synonyms, antonyms etc. (the type of relation becomes visible when you point at the relevant connecting line with the cursor), but also on word based information such as word meaning (*US liquid unit*, *plant organ*) and synonyms of meanings (of the noun: *cupful*, *hole*, *punch*, *incurvature*, as well as of the verb: *form*, *shape*, *transfuse*, *enclose* etc.). The idea is to help users find the right word, either in a text producing situation or in a learning situation, by showing both words and meanings that are related to the central lexical unit. According to the developers, the interface "works like the brain" and allows users to associate intuitively by using an interactive function: if the user clicks on a lexical unit, it is brought to the centre and new words and sense relations for that word appear. In this way, The Visual Thesaurus represents a development which is interesting for us to pursue in a future version as it meets a genuine user need (i.e. finding or learning the right word) through a combination of wordnet and dictionary data, data which are readily available in our resource.

Another approach which we find appealing is the thesaurus-like presentation used by some learner's dictionaries. Figure 7 shows the result of a thesaurus search in

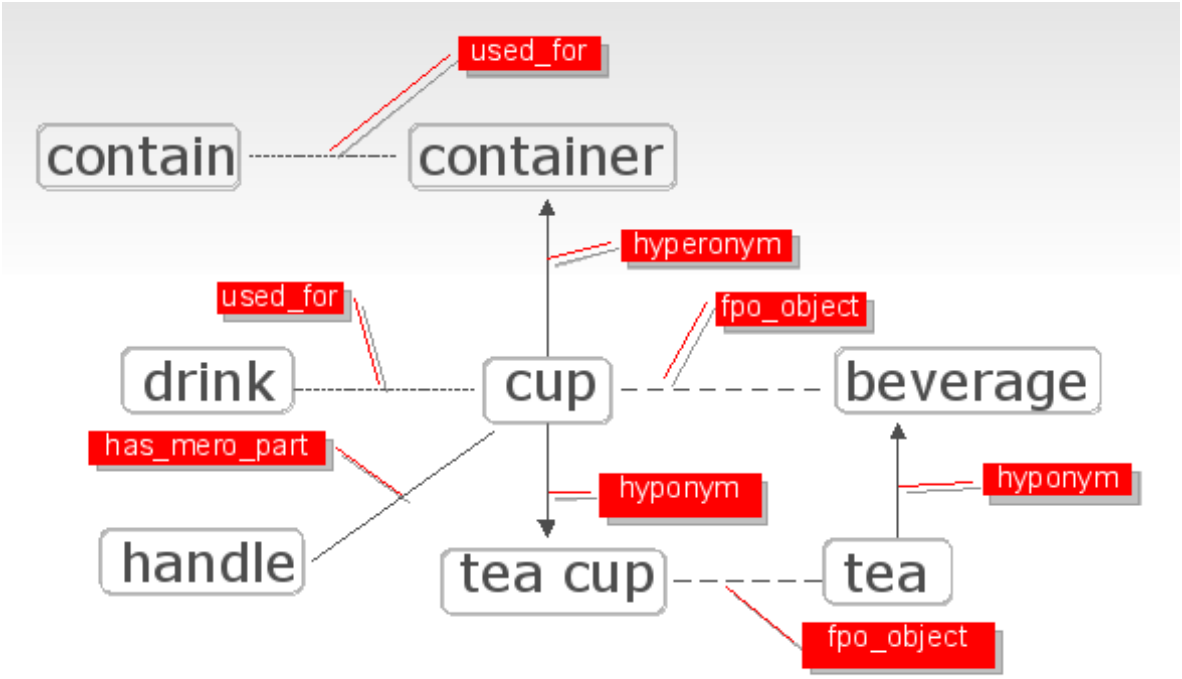

Abbildung 5: A visual presentation of the concept ‘cup’ showing DanNet relations

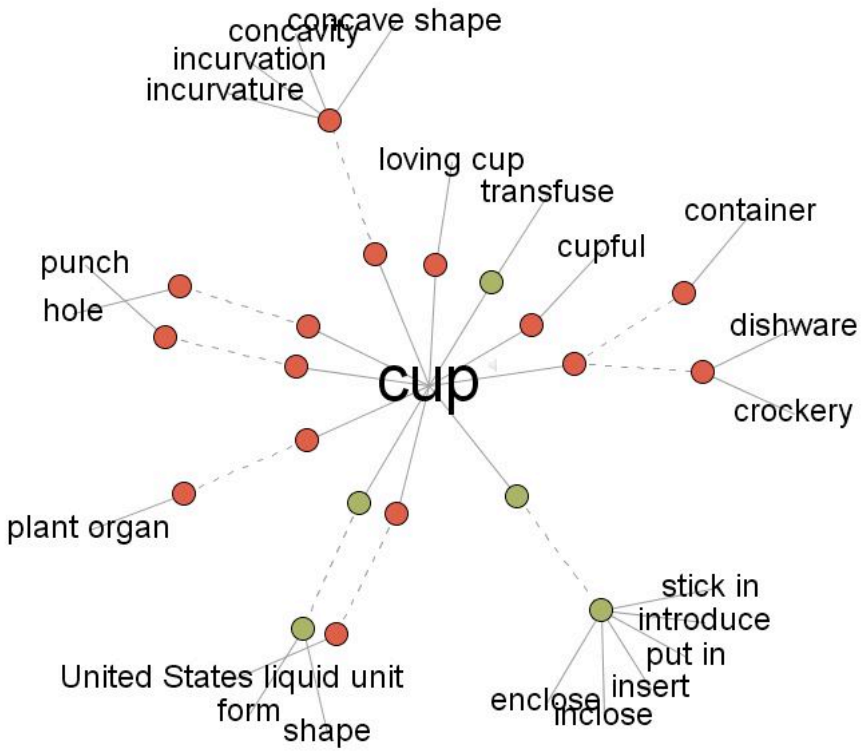

Abbildung 6: The result for ‘cup’ in The Visual Thesaurus

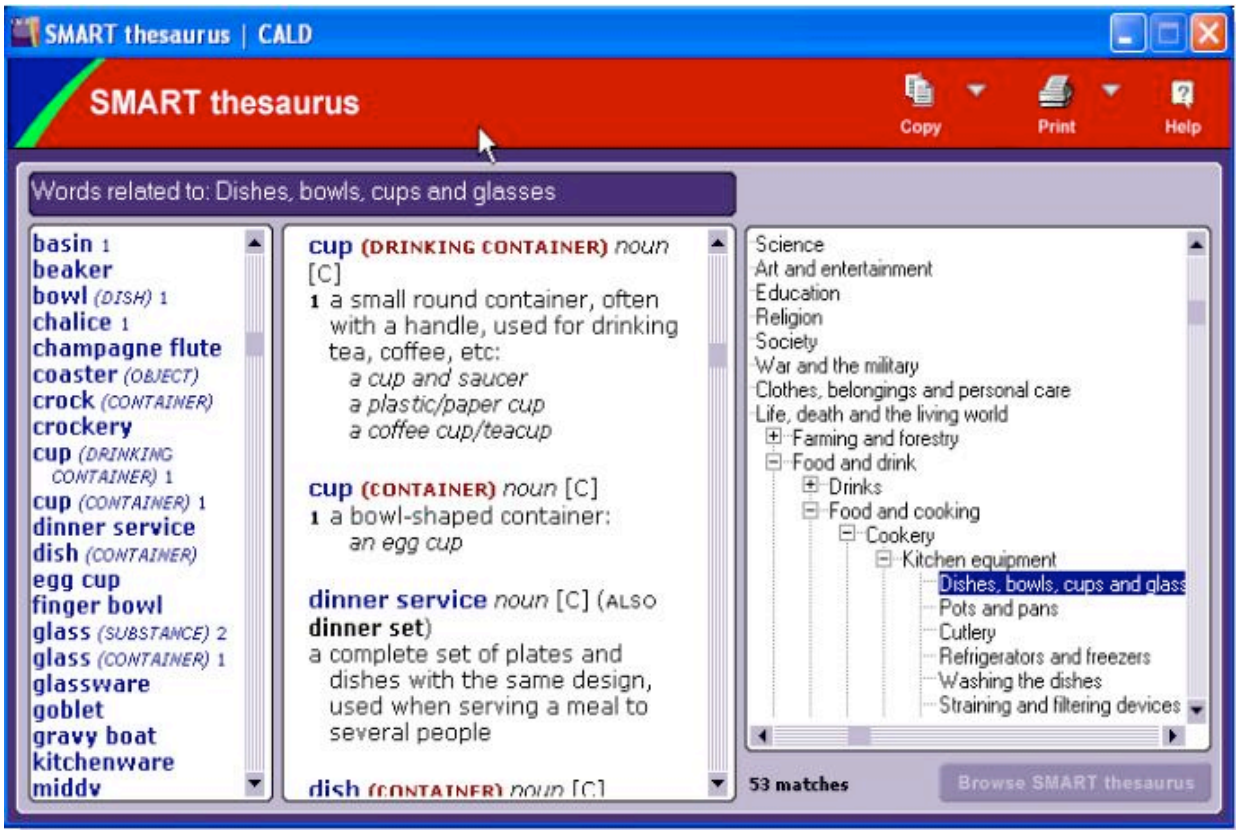

Abbildung 7: The result for 'cup' in the thesaurus of the Cambridge Advanced Learner's Dictionary (CALD)

the Cambridge Advanced Learner's Dictionary (available on CD-ROM or to online subscribers).

Like The Visual Thesaurus, the CALD thesaurus brings together dictionary articles and some kind of conceptual division of the vocabulary, the exact nature of which is not, however, entirely transparent to us. The thesaurus part is shown in the right column and although it might well have used wordnet data either as a point of departure or as a reference point, it seems to have been tailored to a structure that resembles that of Roget's original thesaurus rather than that of a wordnet. The user can navigate through the hierarchy and click to open relevant subcategories until the destination has been found. If the user clicks on a terminal group, the members of that category are displayed in the left column. A click on one of the members opens the corresponding dictionary entry in the central column. In contrast to The Visual Thesaurus, the CALD thesaurus only operates with a single hierarchical dimension based on hypernyms/hyponyms and co-hyponyms. The conceptual categories of the CALD thesaurus seem well suited for human users, in fact often more so than a wordnet hierarchy, in particular in the case where a wordnet category contains a large number of members. Here, the CALD thesaurus categories are more adequately sized and thus easier for human users to grasp. The DanNet hyponym hierarchy could be arranged in a similar way, but not without a substantial amount of manual effort. But as pointed out earlier, a viable semi-automatic

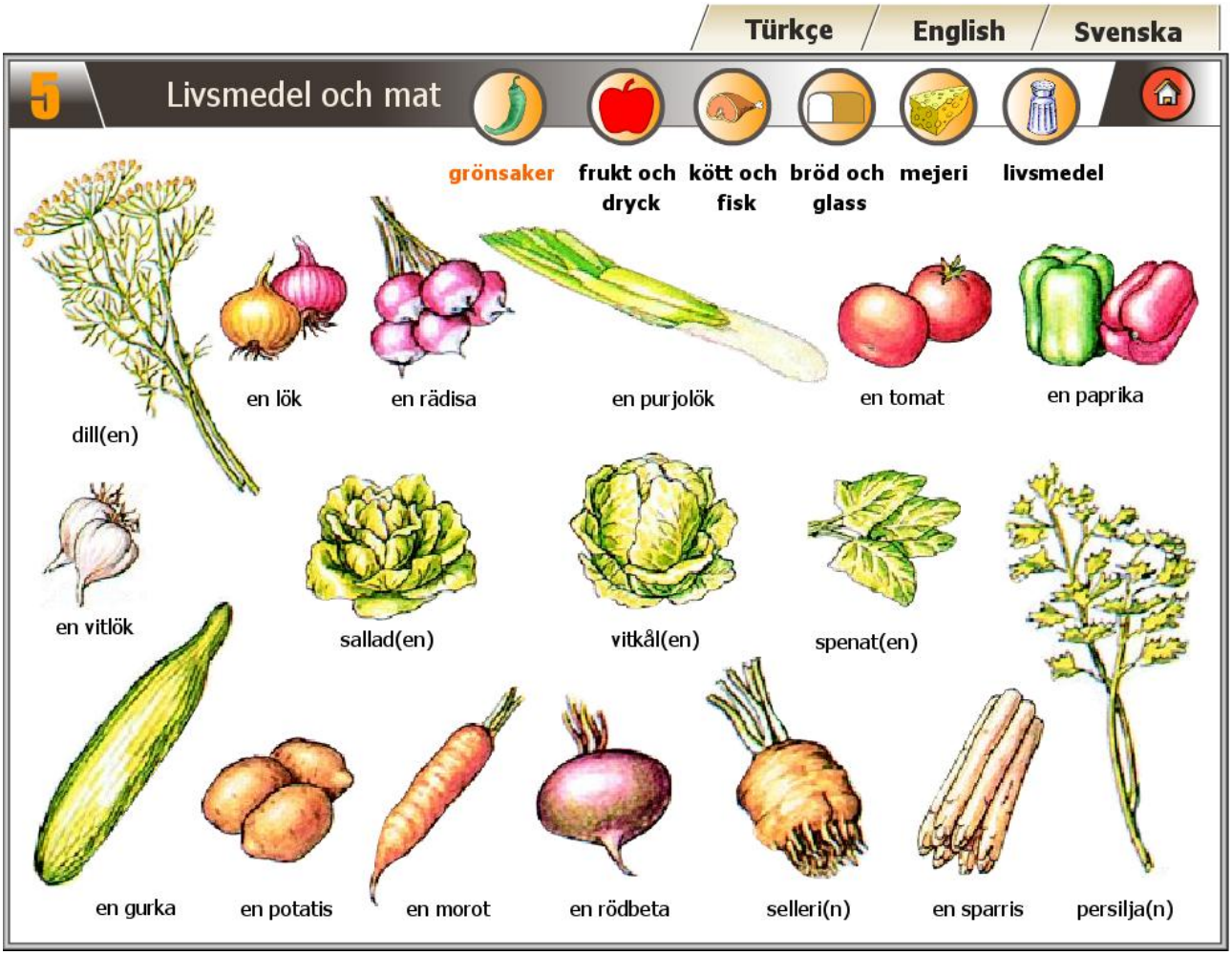

Abbildung 8: Pictures exemplifying the semantic field foodstuffs, taken from the Swedish immigrant dictionary LEXIN

procedure would be to divide large categories according to their ontological type or a relevant relation. Especially in a learning perspective, the presentation of related words belonging to the same domain is useful. For that reason, selected areas are often further developed in learner's dictionaries, and for example supplied with pictures that allow systematic training of vocabulary items (e.g. fruits and vegetables, motor vehicles, kitchen utensils). Figure 8 shows an example taken from the Swedish immigrant's dictionary Lexin.

One final perspective that we would like to mention falls within the area of lexicotainment. Although clearly not among the core functions of an online dictionary, one should not underestimate the role of gadgets and catchy features when it comes to attracting new users or catching the attention of the chance passer-by. Our suggestion is that wordnet data are well suited to improve crossword help in dictionaries. Many online dictionaries have this feature, but all existing helpers that we have come across suffer from the same shortcoming: they ask the user to specify the length of the word

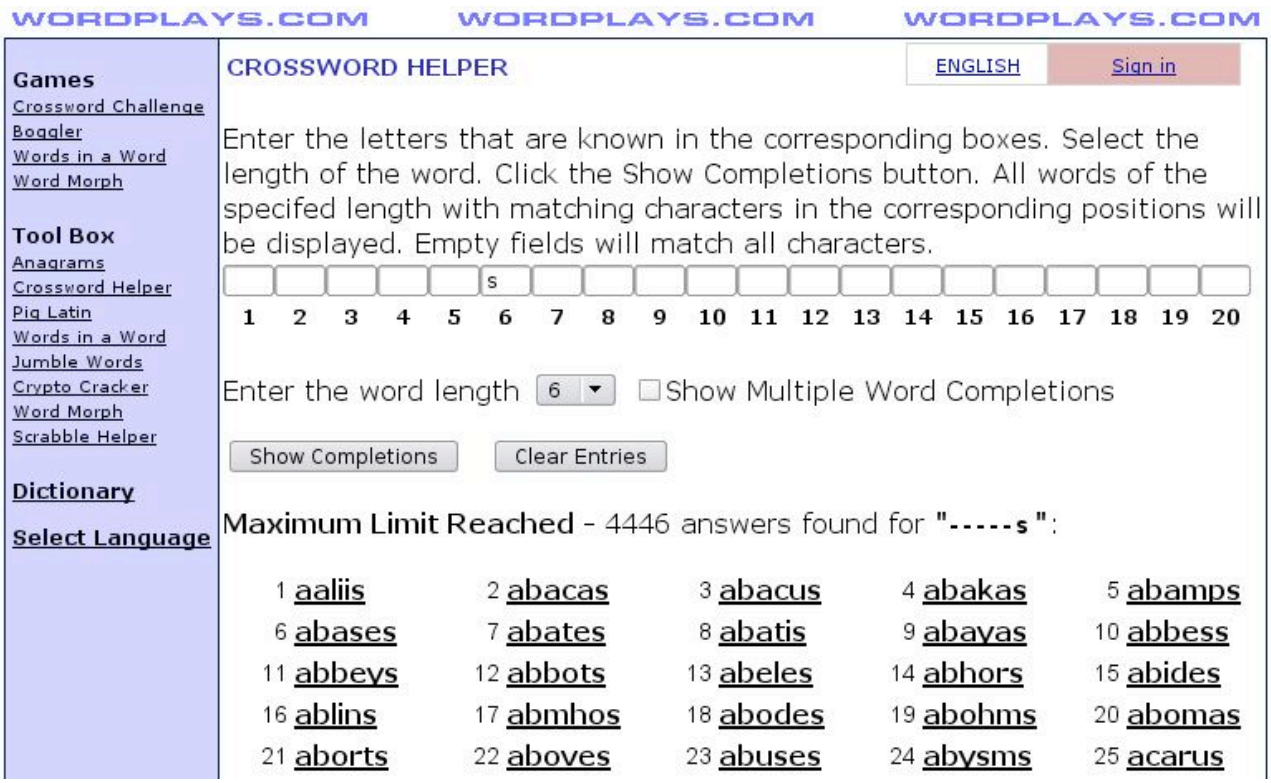

Abbildung 9: A search for six-letter words ending in -s returns 4446 matches

and enter the letters that are known. Subsequently, the input is matched against all words in the database and the results returned. An example is shown in Figure 9 .

The problem is, as anyone who has ever done crosswords will know, that a query based on the string alone overlooks one important factor entirely: the clue. As a result, the string-based crossword helper overgenerates the number of matches and in particular so when the number of known letters are few. Figure 9 shows an example of poor help: the correct matches are lost in the crowd because of the vast number of results. Our suggestion is that dictionary and wordnet data would improve the crossword helper by filtering away undesired matches. This can be done by incorporating information about the clue in at least two respects: using the dictionary's stock of inflectional information would allow the user to search for a particular inflectional form, e.g. a plural noun or a past tense verb, if the clue is in this form. And using wordnet data would allow the user to specify the semantic field within which the query should be performed. In the example above, if the clue was 'sport', a query within this domain in DanNet would quickly provide the answer: *tennis*, *isdans* ('ice dance') and *diskos* ('discus', regarded as a discipline in athletics) are the only three candidates that match.

Literatur

Asmussen, J., Pedersen, B. S., and Trap-Jensen, L. (2007). DanNet – From Dictionary to Wordnet. In Kunze, C., Lemnitzer, L., and Osswald, R., editors, *GLDV-2007 Workshop*

on Lexical-Semantic and Ontological Resources, number 336–3/2007 in Informatik-Berichte, pages 1–9, Hagen. FernUniversität Hagen, Fakultät für Mathematik und Informatik.

DDO (2003–2005). *Den Danske Ordbog 1–6*. DSL & Gyldendal, København.

Lenci, A., Bel, N., Busa, F., Calzolari, N., Gola, E., Monachini, M., Ogonowski, A., Peters, I., Peters, W., Ruimy, N., Villages, M., and Zampolli, A. (2000). SIMPLE – A General Framework for the Development of Multilingual Lexicons. *International Journal of Lexicography*, 13:249–263.

ODS (1956). *Ordbog over det danske Sprog 1–28*. DSL & Gyldendal, København.

Pedersen, B. S. and Nimb, S. (2008). Event Hierarchies in DanNet. In Tanáes, A., Csendes, D., Vineze, V., Fellbaum, C., and Vossen, P., editors, *Proceedings of the Fourth Global WordNet Conference*, pages 339–348, Szeged. University of Szeged, Department of Informatics.

Pedersen, B. S., Nimb, S., Asmussen, J., Sørensen, N. H., Trap-Jensen, L., and Lorentzen, H. (2009). DanNet – The challenge of compiling a WordNet for Danish by reusing a monolingual dictionary. *Language Resources and Evaluation*, to appear.

Pedersen, B. S. and Paggio, P. (2004). The Danish SIMPLE Lexicon and its Application in Content-based Querying. *Nordic Journal of Linguistics*, 27(1):97–127.

Pedersen, B. S. and Sørensen, N. H. (2006). Towards Sounder Taxonomies in WordNets. In *Proceedings from the OntoLex Workshop in association with LREC 2006*, Genova.

Pustejovsky, J. (1995). *The Generative Lexicon*. Bradford Book.

Svensén, B. (1993). *Practical Lexicography: Principles and Methods of Dictionary-Making*. Oxford University Press, Oxford, Translated by John Sykes and Kerstin Schofield edition.

Veale, T. and Hao, Y. (2008). Enriching WordNet with Folk Knowledge and Stereotypes. In Tanáes, A., Csendes, D., Vineze, V., Fellbaum, C., and Vossen, P., editors, *Proceedings of the Fourth Global WordNet Conference*, pages 453–461, Szeged. University of Szeged, Department of Informatics.

Fabienne Martin, Dennis Spohr, Achim Stein

Representing a Resource of Formal Lexical-Semantic Descriptions in the Web Ontology Language

This paper presents an approach to disambiguating verb senses which differ wrt. the inferences they allow. It combines standard ontological tools and formalisms with in-depth formal semantic analysis and is therefore more formalised and more detailed than existing lexical semantic resources like WordNet and FrameNet. The resource presented here implements formal semantic description of verbs in the Web Ontology Language OWL and exploits its reasoning potential based on Description Logics for the disambiguation of verbs in context. After a thorough discussion of the theoretical motivation as well as the manual formal semantic analysis, we present details on the disambiguation process, which is based on a mapping from the French version of EuroWordNet to the Suggested Upper Merged Ontology. In addition to this, we focus on the selectional restrictions of verbs wrt. the ontological type of their arguments, as well as their representation as necessary and sufficient conditions in the ontology. Finally, we discuss how we make use of the Semantic Web Rule Language SWRL in order to calculate the inferences that are permitted on the selected interpretation.

1 Introduction

Verbs raise a number of challenges for computational linguistic applications, two of which will be addressed in this paper. First, a lot of them are highly polysemous, which makes a careful disambiguation a prerequisite for the application of semantic web technologies. As an example, the French verbs *encourager* and *pousser* are normally translated by different verbs in German, as illustrated for *encourager* in (1)-(3), and for *pousser* in greater detail in section 2:

(1) Un terroriste a encouragé$_1$ ma voisine à poser une bombe dans la cave.
A terrorist has encouraged my neighbour to place a bomb in the basement.
Ein Terrorist hat meine Nachbarin ermutigt, eine Bombe im Keller zu legen.

(2) La lettre a encouragé$_2$ ma voisine à poser une bombe dans la cave.
The letter has encouraged my neighbour to place a bomb in the basement.
Der Brief hat meine Nachbarin dazu bewegt, eine Bombe im Keller zu legen.

(3) Le gouvernement a encouragé$_3$ la recherche sur les armes biologiques.
The government has encouraged research on biological weapons.
Die Regierung hat die Erforschung biologischer Waffen angeregt.

Note that (1) differs from (2) only by the ontological category of the subject, which is a human entity in (1) and a non-human one in (2).[1] On the other hand, (3) differs from the previous two in that its object is a non-human entity (while the object denotes a human in (1) and (2)), and in that it does not take an infinitival complement.

The second challenge concerns the computation and the weighting of the inferences triggered by verbs. The contrast between (1) and (2) offers a striking example: while the interpreter of (2) can take for granted that a bomb was placed, she can only *guess* that it was possible in (1). Let us call "actuality entailment" (AE) the entailment triggered by (2) – and to which the interpreter assigns the probability $p = 1$ – that an event satisfying the infinitival complement took place, and "weak inference" the one triggered by (1) – and to which the interpreter assigns a probability p between 0 and 1. Furthermore, we will say that when the AE is triggered, the verb instantiates its "implicative reading", and that it receives its non implicative reading otherwise.[2]

Note that the presence of the non-human subject in (2) is only a necessary condition to trigger the AE (and consequently the implicative reading). The tense of the sentence has also to be of a certain kind (namely a perfective tense, of which the *passé composé* is an example in French) for the AE to arise. The interaction between lexical semantics and information pertaining to the textual level like tense and aspect must then be modelled appropriately to capture the facts.

A model which allows to identify and weigh appropriately the inferences triggered by verbs like *encourager*, *pousser* etc. is highly desirable, since, first of all, verbs of this class are pervasive in the lexicon and heavily present in official texts.[3] Second, rating accurately the inference that an event described by a constituent took place is central for the understanding of texts, or the recognition of textual entailment.

As it is more convenient to present the implementation through specific polysemous verbs, we will first detail the lexical semantics of two specific semantically close verbs at hand, namely *pousser* 'push' and *encourager* 'encourage'.

Plan of the paper. The paper is organised as follows. Section 2 presents the theoretical part of this work by delineating the different readings of the transitive *pousser*, as well as identifying the factors allowing their disambiguation. In addition to this, the semantic differences between *pousser* and *encourager* are discussed. In Section 3, the technical part of the work starts with an introduction to the necessary background. Section 4 discusses the model in detail, focussing on the implementation of selectional argument restrictions as well as the formalisation of inference rules. Section 5 shows how these mechanisms are applied to an example sentence in order to disambiguate

[1] In (2), *encouragé*$_2$ could also be translated figuratively with *ermutigt* instead of *bewegt*.

[2] The terminology is borrowed from Karttunen (1971). Note however that verbs like *encourager* differ from what Karttunen originally calls "implicative verbs" (e.g. *réussir à*, 'manage to'), because the latter trigger an AE with any kind of tenses.

[3] Examples of verbs displaying the alternation between (1) and (2) (and thus triggering inferences of different strength in the two places) are *autoriser à P* 'to authorise to *P*', *inviter à P*, 'to invite to *P*', *aider à P*, 'to help to *P*', *permettre de P*, 'to allow to *P*', *suggérer de P*, 'to suggest to *P*', *exiger que P*, 'to demand that *P*'.

pousser and its syntactic arguments, and to calculate the inferences on the basis of this selection. We conclude in Section 6.

2 Theoretical Background and Motivation: Forceful Verbs in Spatial, Psychological and Conceptual Domains

2.1 Transitive use of *pousser*

2.1.1 Indicators for sense disambiguation

Ontological categories. On its transitive use, *pousser* has roughly four readings, respectively illustrated by the following examples:

(4) Il a poussé Paul vers la porte.
He pushed Paul towards the door. (literal, physical meaning #1, syn. *faire bouger*)

(5) Le pianiste a poussé Fauré vers Brahms.
Lit. *The pianist pushed Fauré towards Brahms.* (figurative, conceptual)
Fig. *The pianist brought the music of Fauré closer to the music of Brahms.*

(6) Il a poussé Paul vers le crime/à tuer.
He pushed Paul towards the crime/to kill. (figurative, psychological)

(7) Il a poussé des cris perçants.
He pushed penetrating cries. (literal, physical #2, syn. *émettre, produire*)

A first obvious indicator allowing to identify the targeted reading is the ontological category of the arguments. The one in (7) is the easiest to disambiguate, since it is automatically selected with an object denoting a sound object (a song, a cry, a spoken word). Disambiguating the other readings is less trivial. First, the physical reading is only available when the subject is able to trigger a physical movement by itself, i.e. is animate. It is for instance excluded in (9), but possible in (8).[4]

(8) Pierre/le vent/la fourmi a poussé y. (physical reading available)
Pierre/the wind/the ant pushed y.

[4] Note that a sentence like *La tristesse a poussé Paul au cimetière*, lit. 'Sadness pushed Paul to the cemetery' is interpreted as describing a physical movement of Pierre, which seems to go against our claim that the physical reading is not available with a inanimate subject. However, we argue that this sentence contains an ellipsis of an à-COMP (*La tristesse a poussé Paul à aller au cimetière*, lit. 'Sadness pushed Paul to go to the cemetery'), which selects the psychological reading of *pousser*. A movement is then interpreted because the AE is triggered with an animate subject and the *passé composé* tense. Two arguments militate in favour of this view. First, the AE disappears with the *imparfait* (the corresponding imperfective sentence does not entail anymore that Paul went to the cemetery), whereas a movement is still entailed with the 'real' physical *pousser*. Second, sentences of this type strongly require the preposition *à* and become deviant with other types of PPs (cf. ??*La tristesse l'a poussé dans le cimetière*, 'Sadness pushed him within the cemetery'), which again suggests that *à* is introducing a hidden infinitive here.

(9) La tristesse/la faim/l'inflation a poussé y. (no physical reading)
The sadness/the hunger/the inflation pushed y.

Second, when the object denotes an abstract entity, the conceptual reading is automatically selected, as in (10). The physical reading is excluded because one cannot physically push an abstract entity, and the psychological one is out because an abstract entity cannot be an Experiencer.

(10) x a poussé ma faim/les prix vers/jusqu'à z. (conceptual reading)
x has pushed my hunger/the prices to the point of z.

When the object denotes a physical entity, all readings are *ceteris paribus* available. The PP then becomes the disambiguating constituent: if the noun contained in the PP denotes an abstract entity, the reading is automatically figurative (psychological or conceptual), cf. (11).

(11) x a poussé y jusqu'à la rage/vers le romantisme. (no physical reading)
x has pushed y to the point of fury/towards romanticism.

To obtain the psychological reading, the object must denote an Experiencer, and the noun contained in the PP must denote a subkind of abstract entities, namely acts or dispositions to act. As e.g. *romantisme* does not pertain to this subset of abstract entities, its presence suffices to select the conceptual reading.

Other indicators. Ontological categories however do not always suffice to isolate the targeted reading, because a lot of nouns are themselves polysemous. The syntactic frame is a further relevant disambiguating indicator. First, the presence of an *à*-COMP suffices to select the psychological reading:

(12) x a poussé y à dormir/à tuer Paul. (psychological reading mandatory)
x pushed y to sleep/to kill Paul.

More interestingly, with an animate subject, the physical readings remain the only ones available in absence of a PP. In other words, with an animate subject, the figurative reading generally makes the PP compulsory. Therefore, leaving out the parenthesised component in the examples in (13) changes the interpretation of *pousser* from figurative to physical. This is also true of other movement verbs like *tirer* 'pull', *jeter* 'throw' or *entrer* 'enter', cf. (13).[5] Note that this fact casts doubt on the idea supported e.g. by Asher and Lascarides (1995) that the syntax remains the same under the metaphorical use of verbs:

(13) a. Le pianiste a poussé Fauré $^{\#}$(vers Brahms). (figurative, conceptual meaning)
Lit.: *The pianist pushed Fauré (towards Brahms).*
Fig.: *The pianist brought (the music of) Fauré closer to (the music of) Brahms.*

[5] To our knowledge, data of this kind are sporadically observed (cf. e.g. Adler and Asnès, 2005), but not explained yet. See Martin (2008) for a theoretical explanation.

b. Albert a poussé Yves #(dans la dépression).(figurative, psychological meaning)
Albert pushed Yves (into depression).
c. Fourier a jeté les étudiants #(dans l'athéisme).
Fourier threw the students (into atheism). (figurative, psychological reading)
d. Pierre entre #(dans l'endettement). (figurative, social reading)
Pierre is entering (into indebtedness).

Apart from some exceptions, the absence of the PPs thus suffices to select the physical reading with polysemous nouns. Therefore, in a sentence like (14), *a poussé* can only be understood as 'physically pushed'.[6]

(14) Le pianiste a poussé le compositeur. (physical reading)
The pianist pushed the composer.

2.1.2 Types of inferences

The inferences triggered by *pousser* on its transitive use basically depend on three factors, namely (i) the ontological type of the subject, (ii) the tense of the verb (cf. Section 1) and (iii) the presence of a PP and the preposition heading it.

On its psychological reading, one of the most relevant inferences concerns the occurrence of the action described in the *à*-COMP or the *à*-object (what we call the AE in the introduction). As already observed above, under its psychological reading, *pousser* entails the occurrence of an action of the Experiencer iff the subject denotes an inanimate entity and the tense used is perfective; cf. (15) as well as (1) to (3).[7]

(15) a. Marie a poussé$_{perf.}$ Paul au suicide/à se suicider.
Marie pushed Paul to commit suicide.
$\not\rightarrow$ Paul is dead.

b. Son divorce a poussé$_{perf.}$ Paul au suicide/à se suicider.
His divorce pushed Paul to commit suicide.
$\rightarrow$ Paul is dead.

c. Son divorce poussait$_{imperf.}$ Paul au suicide/à se suicider.
His divorce was pushing Paul to commit suicide.
$\not\rightarrow$ Paul is dead.

The physical reading also has an 'implicative' and a 'non-implicative' reading, although in another way. As already observed by Jackendoff (1990), *push* does not entail a movement of the Theme ((16) is not contradictory), and the same is true of the French verb *pousser* (cf. Stein (2007) and his example (17)). A movement of the Theme is thus at most a weak inference triggered by *pousser*.

[6] Recall that when the subject is inanimate, the physical reading is not available anyway. The presence of another argument than the subject and the object is thus not necessary to select the figurative reading (cf. e.g. *L'inflation a poussé les prix*, 'the inflation pushed the prices').

[7] Bhatt (1999) and Hacquard (2006) already observed contrasts of the same type for modal verbs, but Mari and Martin (2007) were the first to observe that it expands to non modal verbs like *pousser*.

(16) I pushed the car, but it didn't move.

(17) J'ai poussé la boîte, mais elle n'a pas bougé.
I pushed the box, but it didn't move.

However, in presence of certain PPs (as e.g. *jusque/dans/à*-PPs, but not with *vers/dans la direction de*-PPs), this weak inference becomes an entailment:

(18) J'ai poussé la voiture jusque/dans le garage, #mais elle n'a pas bougé.
I pushed the car until/in the garage, but it didn't move.

(19) J'ai poussé la voiture vers le/dans la direction du garage, *OK* mais elle n'a pas bougé.
I pushed the car towards/in the direction of the garage, but it didn't move.

These contrasts show again that it is crucial (i) to distinguish between strong and weak inferences (entailment or defeasible *a priori* assumptions) and (ii) to differentiate, for a same verb on a same reading, several sets of inferences varying with the syntactic context.

2.2 Semantic representation of transitive *pousser*

The fact that geometrical notions alone do not suffice to model the meaning of all spatial prepositions is already well-documented. For instance, Vandeloise (1991) showed that 'inclusion' cannot alone define the preposition *in*, whose meaning also involves force-dynamic notions like *causation*, *control* or *interaction*. Recently, Zwarts (2007) argued that the same is true of what he calls "forceful verbs" like *push* and *pull*. As he emphasises, spatial notions like *movement*, *direction* and *location* cannot capture the fact that *push* is obviously opposite to *pull*. Indeed, one cannot say that *push* and *pull* describe a movement that goes in opposite direction, since as already noticed above (examples 16 and 17), *push* does not entail a movement of the Theme, and the same is true for *pull*. It is not even clear that these verbs entail a movement of the Agent (arguably, exerting a pressure on something without moving can still be a *pushing/pulling*). According to Zwarts, it is rather the *forces* involved which go in opposite directions: the force is pointing away from the Agent in the case of *push*, and towards him with *pull*.

2.2.1 Vectors

Following Zwarts (2007), we will model the informal notion of force with the help of the mathematical notion of vectors. A vector v is an arrow, i.e. a directed line segment. *Free* vectors have a length and direction, but no starting point. *Located* vectors have a starting point *st-p(v)*. *st-p(v)=x* means that x defines the starting point of the vector v. On the other hand, forces have two parameters, namely a *magnitude* $m(v)$ and a *direction* $dir(v)$. $dir(v) = y$ indicates that the vector v points into the direction of

y. Here, we will slightly augment the formalisation of Zwarts in order to define more precisely the direction of v. $dir(v) = y, z$ indicates that the vector v points into the direction of y and is parallel to the line joining y to z. This allows one to define on which side of y the pressure is exerted.

The two parameters m and d define a vector (the length of the vector co-varies with the magnitude of the force it represents). The *location* $loc(v)$ of the force corresponds to the physical point where the force is exerted. After Talmy (1985) and following Zwarts (2007), we will call *Antagonist* the Agent exerting the force, and *Agonist* the Patient on which the force is exerted.

Objects often tend to move by themselves in a particular direction. In that case, the Agonist has also its own force vector, which represents its inherent tendency to move by itself (generally downwards, because of gravitation). A *resultant vector* determines the result of the interaction between the forces of the Antagonist and the Agonist. This sum can be zero, when the forces of the Antagonist and the Agonist are equal but opposite.

Let us see now how Zwarts defines the meaning of *push* and *pull* with the help of vectors. Informally, their meaning can be represented in the following way. The arrows represent the force vector.

(20) *push*: Antagonist ⇒ Agonist

(21) *pull*: Antagonist ⇐ Agonist

It can be modelled more formally in a vector model. Let us assume that the vector v_{sp} represents the *spatial* relation between the Antagonist and the Agonist pointing from the Agonist to the Antagonist. v_{sp} provides the spatial frame for the $force$ vector v_f. What *push* and *pull* express is the way v_f is aligned with respect to v_{sp}: these two vectors are opposite for *push*, and point in the same direction for *pull*. If no other forces are interacting, the Agonist will move in the direction of the force vector, i.e. away from the Antagonist in the case of *push* (and in its direction in the case of *pull*).

We observe that the definition of Zwarts rightly predicts that *push* is not appropriate when the context indicates that the force vector v_f points into the direction of the Antagonist (cf. (22)) and that *pull* is not appropriate when it indicates v_f points into the opposite direction of the Antagonist (cf. (23)):

(22) #J'ai poussé le landau vers moi.
I pushed the buggy towards me.

(23) #J'ai tiré le landau vers le mur en face de moi.
I pulled the buggy against the wall opposite of me.

2.2.2 Lexical representations

We propose to assume that forceful verbs introduce an implicit vector argument. This argument can be targeted by modifiers like *hard* (cf. *push hard*). Source(v, e) means that the event e is the source of the vector force v. We can then define the two forceful

movements *pousser* 'push' and *tirer* 'pull' in the following way. Recall that *st-p(0)=x* means that x is the starting point of the vector v, $loc(v)$=y that y corresponds to the entity on which the force is exerted, and $dir(v)$=y, z that v is parallel to the line joining y and z. In absence of a spatial complement, z is left underspecified and the interpretation by default given the context is chosen. In presence of such a spatial PP, z is denoted by the noun in the PP.

(24) $pousser_{tr} \rightsquigarrow \lambda z \lambda y \lambda x \lambda e \lambda v[\mathsf{Antagonist}(e, x) \wedge \mathsf{Agonist}(e, y) \wedge \mathsf{Source}(v, e) \wedge \mathsf{st\text{-}p}(v) = x \wedge \mathsf{loc}(v) = y \wedge \mathsf{dir}(v) = y, z \wedge m(v) > 0]$

(25) $tirer \rightsquigarrow \lambda z \lambda y \lambda x \lambda e \lambda v[\mathsf{Antagonist}(e, x) \wedge \mathsf{Agonist}(e, y) \wedge \mathsf{Source}(v, e) \wedge \mathsf{st\text{-}p}(v) = x \wedge \mathsf{loc}(v) = y \wedge \mathsf{dir}(v) = x, z \wedge m(v) > 0]$

It is specified that the magnitude of the force vector must be superior to zero, because a *pushing/pulling* of zero magnitude is not a *pushing/pulling*. Note that this representation is underspecified wrt. the domain (e.g. spatial, psychological, social or conceptual) on which *pousser* or *tirer* are used.

2.2.3 Presupposition

This representation is still too weak though. To see it, let us compare *pousser y* and *glisser y* ('slide y'), which is arguably another forceful verb on its transitive use. A first difference, irrelevant here, is that *glisser y* entails a change of state of y, which explains the contradiction in (26):

(26) Il a glissé la lettre, #mais elle n'a pas bougé.
He slid the letter, but it didn't move.

The important point is that we do not want to say that *glisser y* is equivalent to *pousser y* modulo this entailed change of state. Indeed, while *glisser y* indicates that no other forces interact with the one initiated by the Antagonist to make y move, *pousser y* presupposes such an interacting resistant force. For instance, *pousser la lettre dans la boîte* ('push the letter in the box') will be chosen in a context where the opening of the box is stuffed with something which somehow blocks the insertion of the letter in it. In a context where resistance is totally absent, the use of *pousser* is not appropriate, cf. (27):

(27) Le vent a poussé OK la fumée épaisse qui se dégageait de l'incendie/ #l'air pur du matin.
The wind has pushed the thick smoke which spread from the fire / the clear morning air.

(28) J'ai tiré le livre de la bibliothèque.
I've pulled the book from the bookshelf.

Similarly, (28) is fully appropriate only if the book is somehow blocked in its position on the bookshelf (because it is between two books for instance).[8] On the contrary, *glisser* in *glisser la lettre dans la boîte* ('slide the letter in the box') will be preferred when nothing makes obstacle to the movement initiated by x. The presence or absence of a resistant force also differentiates *pousser/tirer* from *glisser* in the non-spatial domain:

(29) L'interprète a fait glisser Cage vers Satie.
The interpreter let Cage slide towards Satie.

(30) L'interprète a tiré/poussé Cage vers Satie.
The interpreter pulled/pushed Cage towards Satie.

For instance, (30) suggests that some force crosses the one initiated by the interpreter who wants to narrow the conceptual distance between the music of Cage and Satie, while (29) precisely suggests the absence of such a force (*la voie est libre*, 'the coast is clear').[9]

It is pointless to define the interacting vector by its position with regard to the spatial/conceptual vector linking x to y, because it can vary considerably from one context to another. But we can simply state that if v takes place, a vector v' interacting with v and of a non-null magnitude is also present.

Importantly, the entering into play of the interacting vector is not *asserted* by the forceful verbs. It rather seems to be presupposed. Look for instance to the following sentences:

(31) Il est possible qu'il ait poussé la lettre dans la boîte.
It is possible that he had pushed the letter in the box.

(32) L'interprète n'a pas poussé Brahms vers Fauré.
The interpreter hasn't pushed Brahms towards Fauré.

The speaker of (31) clearly assumes that the box entrance is stuffed with some matter which *would* originate in an interacting vector if somebody tried to push something in it, and the speaker of (32) assumes the existence of a conceptual obstacle between Brahms and Fauré which *would* trigger an interacting vector if somebody tried to narrow the two. It confirms the presuppositional nature of the inference, since (31) and (32) are classical presuppositional environments. Note that what is presupposed is not the interacting vector itself, but the obstacle between y and z that could trigger it.

Interestingly, the presupposition is kept under the intransitive use of *pousser*. For instance, when it means *aller* 'to go', the intransitive *pousser* presupposes that the

[8] Intriguingly, no interacting force is implied by ***re**tirer*. For instance, while *J'ai retiré la plume du bureau* ('I withdrew the feather from the desk') is fine, *#j'ai tiré la plume du bureau* ('#I pulled the feather off the desk.') is odd precisely because it is difficult to imagine a force resisting to the one of the Antagonist. Another difference between the two is that *retirer* but not *tirer*, is a hyponym of *enlever*, which explains its incompatibility with directional PPs.

[9] Note that the periphrastic causative in *faire*+inf. is almost compulsory in (29) to trigger the non-physical reading of *glisser*. The lexical causative *L'interprète a glissé Cage vers Satie* oddly suggests that the musician physically pushed Cage towards Satie. We will ignore these data here.

displacement was made difficult by an obstacle manifesting a certain resistance to the Performer of the movement. This explains why (33) is acceptable whereas (34) is strange when uttered out of the blue, since going as far as the kitchen generally does not consist in a big achievement:

(33) Je suis allée jusqu'à la cuisine.
I went until the kitchen.

(34) #J'ai poussé jusqu'à la cuisine.
Lit.:*I pushed until the kitchen.*
Int.:*I went until the kitchen.*

2.3 Comparison between *pousser* and *encourager*

Obviously, it is on the psychological reading that *pousser* resembles more *encourager*. The analysis presented above allows us to easily capture the differences between the two near-synonyms. First, *encourager* is not classified as a movement verb, and as such is not acceptable with strictly directional PPs:

(35) *Pierre m'a encouragé vers le crime.
Pierre encouraged me towards the crime.

Second, *encourager* does not trigger the presupposition described above. This difference is responsible for the following contrast[10]:

(36) L'entraîneur a encouragé/?poussé l'équipe à gagner.
The trainer has encouraged/pushed the team to win.

(37) L'entraîneur a encouragé/?poussé l'équipe à bien jouer.
The trainer has encouraged/pushed the team to play well.

(38) Il a bien voulu m'encourager/?me pousser à le faire.
He has been so kind as to encourage/push me to it do.

In the psychological domain, the occurrence of an interacting vector presupposed by *pousser* translates in the following way: it indicates that x and y entertain inverse preferences; x wants y to do P (P corresponding to the proposition denoted by the infinitive), while the speaker or y prefers $\neg P$ (the 'conflict' between the two preferences can be conceived as the psychological translation of the 'resistance' between the Pusher and another entity). This explains why *pousser* becomes deviant when x and y normally entertain the same preferences.

[10] Of course, the contrast is only present if we do not assume that the team wants to lose. In any case, it is safe to say that the resistance of the team seems to be greater with *pousser* than with *encourager*.

3 Technical Background

Since the theoretical background has been established, we will now turn to the more technical part of this work. In this section, we will briefly outline the technical background, i.e. the main features of the formalisms as well as the lexical-semantic and ontological resources used in this study.

3.1 Formalisms

The formalisms that we use for building our resource have been developed in the field of the *Semantic Web*, a research area devoted among others to providing tools and formalisms for assigning meaning to web content (Berners-Lee et al., 2001). In particular, we make use of the Web Ontology Language OWL (Bechhofer et al., 2004) and the Semantic Web Rule Language SWRL (Horrocks et al., 2004).

3.1.1 OWL

The Web Ontology Language (Bechhofer et al., 2004) is a formalism based on the Resource Description Framework RDF[11] and can be expressed in XML syntax. Its main building blocks are classes (corresponding to one-place predicates in first-order logic) and properties (two-place predicates), which are both structured hierarchically and inherit restrictions (such as axiomatic definitions of classes or, in the case of properties, formal characteristics like being functional or transitive) to their subclasses and subproperties respectively. In addition to these entities, there are also individuals, which are simply instances of classes. OWL comes in three sublanguages, which differ wrt. their expressivity: OWL Lite is the least expressive sublanguage and allows for simple class definitions; OWL DL is based on Description Logics (DL), a decidable fragment of first-order logic (see e.g. Baader et al., 2003), which restricts the use of some OWL constructs in order to maintain decidability of reasoning; OWL Full is the most expressive sublanguage and imposes no restrictions on the language constructs, however at the cost of decidability. For example, in OWL Full it is possible to express that a class is an instance of another class, which is disallowed in OWL DL.

Resources defined in OWL Lite and OWL DL can be interfaced with a Description Logic reasoner (e.g. Pellet; Sirin et al., 2007) on the one hand to check the consistency of the resource, and on the other hand to infer further statements on the basis of explicit statements in the resource. Such reasoning tasks are e.g. the classification of the taxonomy in order to infer logical subclass relationships, or the classification of individuals for inferring the classes they implicitly instantiate, based on necessary and sufficient conditions in the axiomatic definitions of classes.

[11] `http://www.w3.org/RDF/`

3.1.2 SWRL

The Semantic Web Rule Language (Horrocks et al., 2004) adds expressivity to OWL in that it allows for the expression of Horn-like rules, i.e. disjunctive rules with at most one positive literal, as in the two equivalent formulae in 39 and 40.

(39) $\neg hasFather(x, y) \vee \neg hasBrother(y, z) \vee hasUncle(x, z)$ (Horn clause)

(40) $hasFather(x, y) \wedge hasBrother(y, z) \rightarrow hasUncle(x, z)$ (equivalent SWRL rule)

SWRL can be expressed directly in OWL syntax – so the resulting documents are still OWL compliant – and the rules can be interpreted and executed by tools such as the Jess® rule engine[12].

3.2 Lexical-semantic and ontological resources

3.2.1 EuroWordNet

The EuroWordNet project (Vossen, 1998) aimed at providing resources similar to Princeton WordNet (Fellbaum, 1998) for eight European languages, all of which are connected through an interlingual index (ILI) that contains a set of language-independent concepts. The ILI is linked to the so-called EuroWordNet Top Ontology, an upper-ontology-like collection of features that have been designed to describe the lexical-semantic relations in the wordnet. The French version of EuroWordNet contains roughly 8,300 verb senses and 24,500 noun senses, which are organised into 22,745 synonym sets and linked using lexical-semantic relations like hyponymy and meronymy.

In contrast to the scale of the resource in terms of covered senses, the detail of description is generally limited to taxonomic relations between synonym sets and does not include information on argument structure. However, the probably biggest drawback of the French EuroWordNet lies in its inaccuracy and even partial incorrectness, mainly wrt. the verbal descriptions. Therefore, only the noun hierarchy can be considered as a useful starting point for building other lexical resources, whereas the verb hierarchy can only provide a rough sketch as to the interpretation and organisation of the senses.

Other resources. Apart from EuroWordNet, there is no large-scale lexical resource of French that provides qualitatively adequate lexical-semantic analyses. While resources such as FrameNet and VerbNet (Baker et al., 1998; Kipper-Schuler, 2006) exist for English, none of these have been extended to French in a comparable way so far.

3.2.2 SUMO

Together with DOLCE (see below), the Suggested Upper Merged Ontology (Niles and Pease, 2001) is one of the most widely used ones in the NLP community, among others due to the fact that mappings have been created to Princeton WordNet (Niles and

[12] http://www.jessrules.com/

Pease, 2003) and the EuroWordNet ILI (Spohr, 2008a). SUMO comes with MILO, a mid-level ontology, as well as domain ontology extensions, which in total contain 20,000 terms and 70,000 axioms. While originally implemented in SUO-KIF – a formalism intended as first-order language – SUMO has also been translated to OWL Full, with the attempt to preserve as much as possible of the original axiomatisation.

Despite its quantitative size and degree of formalisation, SUMO has been criticised primarily wrt. the usability of its axiomatisations, since they are questionable from a modelling perspective (e.g. instances being concepts at the same time and relations being modelled as concepts). Moreover, SUMO seems to lack a clear theoretical basis, as it adopts ideas from different ontological theories (Sonntag et al., 2007).

3.2.3 DOLCE

The Descriptive Ontology for Linguistic and Cognitive Engineering is an upper-level ontology that has been designed with a strong cognitive bias (Gangemi et al., 2003a). Its classes and the relations among them have been implemented with the OntoClean methodology (Guarino and Welty, 2002), which gives the resource a formally and theoretically more solid basis than e.g. SUMO. As was mentioned above, DOLCE has also been mapped to Princeton WordNet (Gangemi et al., 2003b).

DOLCE is the first reference module of the WonderWeb library of foundational ontologies, and it has a number of extensions (e.g. an ontology of information objects). In total, DOLCE and its extensions comprise roughly 200 classes and 300 properties, and they are available as OWL versions.

4 Implementation of the Model

In the following, we will describe how we model the inference triggers mentioned in Section 2 as well as formal semantic representations like the one depicted in (24) using OWL DL and the more expressive SWRL.

4.1 Encoding of selectional argument restrictions

As was mentioned in the introduction, the primary triggers for selecting one particular meaning over another is the presence (or absence) of syntactic arguments as well as their ontological type. For example in (2), the fact that (i) *encourager* subcategorises an infinitive, (ii) the subject is inanimate, and (iii) the direct object is animate, determine the sense of *encourager* in this sentence. In order to make this information available and processable, we use a straightforward encoding of these triggers as conditions on class definitions (see also the representations in Franconi (2003)), based on conceptual classes of the DOLCE-Lite-Plus ontology and its extensions (in the examples below prefixed by *dol:* for DOLCE and *edns:* for "Extended Descriptions and Situation").

In particular, the different senses of a verb are modelled as subclasses of a general class that denotes an underspecified representation of the verb. However, the different

verb senses are only subclasses of this generic representation in the lexicon, not in our concept hierarchy, since verb senses very frequently denote different concepts that are not subsumed by a common concept. (41) below shows the class definition of the conceptual sense of *pousser* that corresponds to the one used e.g. in sentence (5) above.

(41) $pousser\_conceptual \equiv pousser$
$\exists\ subj\ (\exists\ canDenote\ edns{:}agentive\text{-}social\text{-}object)$
$\exists\ obj\ (\exists\ canDenote\ dol{:}abstract)$
$\geq\ 3\ arg\ owl{:}Thing$

The formalisation is to be interpreted as follows: in order to be classified as an instance of *pousser_conceptual*, it is both necessary and sufficient to be an instance of *pousser*, with a subject that can denote an agentive social object, with a direct object that can denote something abstract, and with at least one more argument (i.e. the number of values for *arg* – which is the superproperty of *subj*, *obj* and further argument properties – is at least 3; *owl:Thing* just refers to "any kind of entity"). The predicate *canDenote* used in the formalisation captures the polysemy of the nominal argument, since the classes that represent nouns contain as axioms the ontological concepts they can denote, such as e.g. the class *compositeur* with the axiom ∃ *canDenote edns:agentive-social-object*. So in other words, the subject part of the example above states that the value of the *subj* property of pousser has to be an instance of a class that can denote an agentive social object.

The properties *subj* and *obj* have been defined as *functional properties*, i.e. they can only have one value. Thus it suffices to use the existential quantifier in the axiomatisation here. However, in other cases it is necessary to use a combination of the existential and the universal quantifier, e.g. in order to express that for *pousser_physical* – irrespective of the particular syntactic configuration – it is necessary that all arguments denote something that is not abstract. Conversely, it would not suffice to use just the universal quantifier here, since that would be trivially satisfied in a case where *pousser* is used without any arguments. Thus, a definition of *pousser_physical* has to contain at least the following statements.

(42) $pousser\_physical \equiv pousser$
$\exists\ arg\ (\exists\ canDenote\ \neg dol{:}abstract)$
$\forall\ arg\ (\exists\ canDenote\ \neg dol{:}abstract)$
$\ldots$

The general motivation for the encoding shown in (41) and (42), which views the contextual triggers discussed above as necessary and sufficient conditions, is that a DL reasoner can infer – on the basis of a particular setting of contextual parameters (i.e. property values) – the specific type of an instance of the generic *pousser*. In the following, we will discuss the inference rules that are attached to each sense class, and which are evoked once a specific sense has been determined.

4.2 Inference rules

As was mentioned above, the different senses of *pousser* do not only differ wrt. necessary and sufficient conditions that are used to classify them, but also wrt. the inferences that may be drawn from them. In our resources, such inferences are encoded in the form of SWRL rules (see e.g. O'Connor et al., 2005), as they require inference capacities which go beyond the scope of the inventory provided by OWL DL. In order to keep the following discussion as simple as possible, we will restrict ourselves to explaining the inference rule that corresponds to the semantic description of transitive *pousser* given in (24). The SWRL rule is shown in Table 1 below, with the rule body in lines 1 to 6 and the rule head in lines 7 to 19.

The first line represents the configuration in which the rule is applicable, i.e. an instance of *pousser* with grammatical subject and object. Lines 2 to 6 make use of the SWRL extensions built-ins[13] defined within the Protégé ontology editor (Knublauch et al., 2004) in order to create the instances that are to be inserted into the representation, based on the description in (24). In line 7, a *PUSHING* event is asserted. In lines 8 and 9, the grammatical subject and object are asserted as the antagonist and agonist of the event denoted by *pousser*. Lines 10 and 11 assert a vector which has as its source the *PUSHING* event. Lines 12 to 15 assert locations which correspond to the location of the grammatical subject, the grammatical object and the underspecified entity z respectively (cf. page 7). In addition to this, the location of the subject is further the starting point of the vector (16), the location of the object is the location of the vector (17), i.e. the location where the force is exerted, and line 18 specifies that the direction of the force of the vector is parallel to the line that joins the grammatical object and the underspecified entity z, i.e. parallel to a line that has the location of y and z as points. Finally, line 19 states that the magnitude of the vector is 1.

5 Disambiguation and Calculation of Inferences

In order to select the correct reading of a verbal predicate in a sentence like (14) and, moreover, to generate the appropriate semantic representation on the basis of this choice, our system passes a number of distinct analysis steps. Basically, the system receives input from a syntactic parser and tries to determine the correct senses of both the verbal predicate and its syntactic arguments, before calculating the inferences permitted on this interpretation. The whole analysis process is summarised in Figure 1 below.

For the scope of this paper, we will ignore details on the syntactic analysis that precedes the semantic processing steps, and instead assume a syntactic parser which returns output like the one depicted in Figure 1, providing information on the predicate (*pousser*), its syntactic arguments (*pianiste* and *compositeur*), its modal context (e.g. embedding under *pouvoir*, 'can, be able to'), and the tense in which the predicate is

[13] See `http://protege.cim3.net/cgi-bin/wiki.pl?SWRLExtensionsBuiltIns`; the built-in function `createOWLThing` has been replaced with `cOT` in the table. One could say that `createOWLThing` represents the existential quantifier, although its interpretation is somewhat stronger since actual instances are asserted and created in the resource.

1	syntactic configuration required for application of rule	`pousser(?e) ∧ subj(?e,?x) ∧ obj(?e,?y) ∧`
2	create vector for representing the force	`swrlx:cOT(?v,?e) ∧`
3	create spatial location of grammatical subject	`swrlx:cOT(?locx,?x) ∧`
4	create spatial location of grammatical object	`swrlx:cOT(?locy,?y) ∧`
5	create underspecified entity	`swrlx:cOT(?z,?e) ∧`
6	create spatial location of underspecified entity	`swrlx:cOT(?locz,?z) ∧`
		→
7	assert pushing event	`PUSHING(?e) ∧`
8	assert grammatical subject as antagonist of the pushing	`antagonist(?e,?x) ∧`
9	assert grammatical object as agonist of the pushing	`agonist(?e,?y) ∧`
10	assert a vector	`VECTOR(?v) ∧`
11	assert the pushing event as source of the vector	`source(?v,?e) ∧`
12	assert locations	`LOCATION(?locx) ∧ LOCATION(?locy) ∧ LOCATION(?locz) ∧`
13	assert location of the grammatical subject	`hasLocation(?x,?locx) ∧`
14	assert location of the grammatical object	`hasLocation(?y,?locy) ∧`
15	assert location of the grammatical object	`hasLocation(?z,?locz) ∧`
16	assert the location of the grammatical subject as the starting point of the vector	`hasStartingPoint(?v,?locx) ∧`
17	assert the location of the grammatical object as the location of the vector	`hasLocation(?v,?locy) ∧`
18	assert the direction of the force of the vector to be parallel to the line joining the antagonist and the agonist (cf. page 7)	`LINE(?l1) ∧ LINE(?l2) ∧ hasPoint(?l1,?locy) ∧ hasPoint(?l1,?locz) ∧ parallel(?l1,?l2) ∧ hasDirection(?v,?l2) ∧`
19	assert the magnitude of the vector	`hasMagnitude(?v,1) ∧`

Table 1: SWRL rule corresponding to the semantic description of *pousser* in (24)

used[14]. These context features are crucial for determining the inferences that may be drawn, and thus play an important role in the semantic processing steps which build on the syntactic analysis (see below).

Disambiguation of the predicate and its syntactic arguments. In order to select the correct sense of the verbal predicate, we first disambiguate its syntactic arguments. For this, we apply a similar methodology to that presented in Spohr (2008b). In essence, this approach makes use of a mapping between the French EuroWordNet – a lexical semantic resource for French (EWN; see e.g. Vossen (1998)) – and SUMO (Niles and Pease, 2003), and that has recently been extended to provide a mapping to DOLCE.

[14] For a more detailed discussion of the syntactic analysis, the reader is referred to Boullier and Sagot (2005).

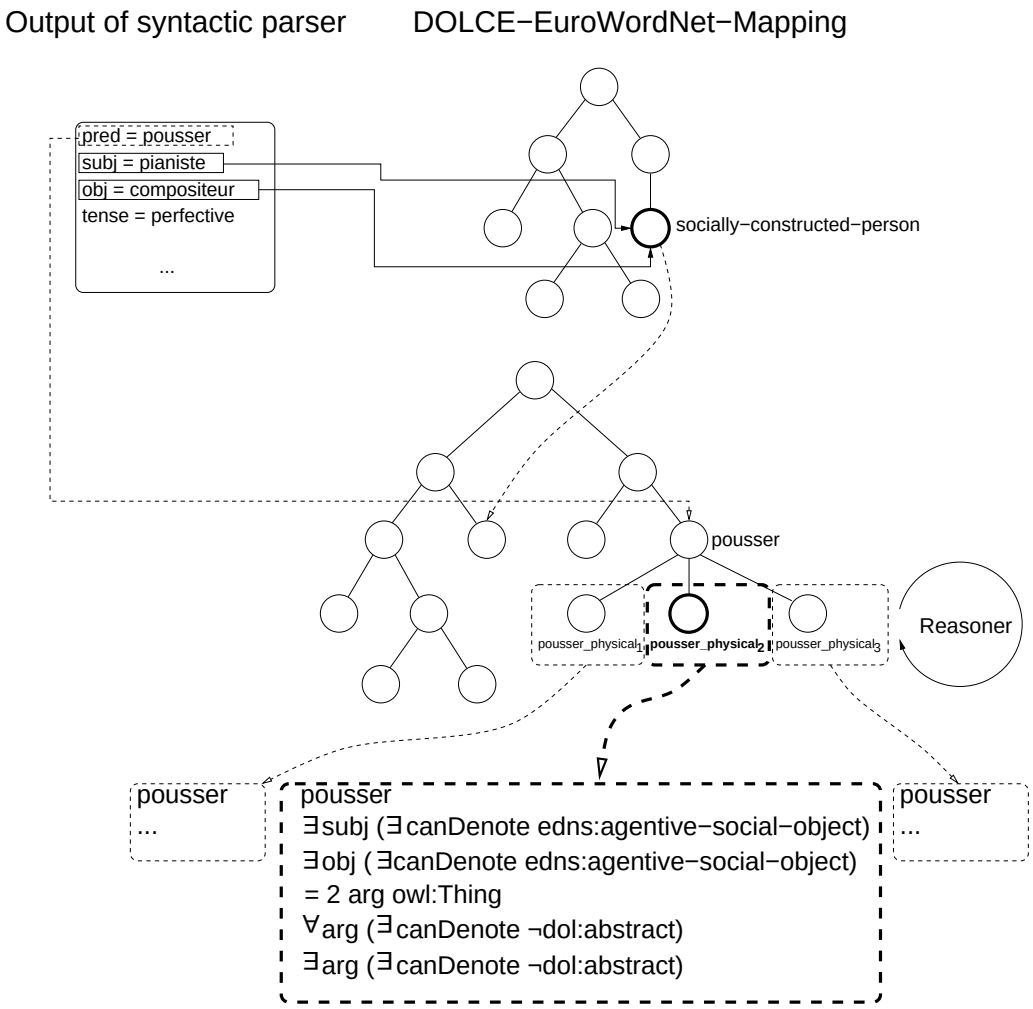

Figure 1: Schema of the process of determining the intended sense of $pousser$ in (14) from syntactically parsed input

On the basis of this mapping, selectional preferences are calculated and expressed in terms of ontological concepts, rather than EuroWordNet synsets. Thus, by applying this methodology to a verb like *pousser*, we obtain lists of selectional preferences wrt. the ontological types of its subject and object (see top righthand corner of Figure 1). For the actual disambiguation, the different senses of the subject (*pianiste* in the present case) are looked up in the DOLCE-EWN mapping, and the sense scoring highest in the corresponding selectional preference list is selected. The words are then asserted as instances of the respective EWN classes (in this case *pianiste_1* and *compositeur_2*), which in this case have the necessary condition ∃ *canDenote soc:socially-constructed-person*, a subclass of *edns:agentive-social-object* (cf. (41) above).

The output of the process of disambiguating the arguments is, of course, not entirely deterministic. However, when viewed from the highly abstract level of ontological concepts, the senses distinguished in EWN are very often still closely related so that their sense distinctions have no impact on the interpretation of the verbal predicate and thus the selection of the appropriate sense. For example, although there are two senses of *compositeur* distinguished in EWN – the "non-musical" *compositeur* being that in the sense of a typographer –, they are still subsumed under the common DOLCE class *socially-constructed-person*, which suffices to select the correct sense of *pousser* irrespective of the particular interpretation of *compositeur*. Therefore, even though some of the arguments may be disambiguated towards the wrong sense, the interpretation of the verb sense stays the same and thus the inferences drawn on the basis of this selection remain unaffected. Therefore, even though some of the arguments may be disambiguated towards the wrong sense, the interpretation of the verb sense stays the same and thus the inferences drawn on the basis of this selection remain unaffected.

Once the syntactic arguments have been disambiguated, they are linked to the instance representing *pousser*. The intermediate representation obtained from the operations so far looks as follows.

(43) $pousser(e) \wedge subj(e, pianiste) \wedge obj(e, compositeur) \wedge pianiste\_1(pianiste) \wedge compositeur\_2(compositeur)$

The next step consists in determining the correct sense of *pousser*. As was mentioned in Section 4 above, selectional restrictions have been implemented as necessary and sufficient conditions on class definitions, which allows a reasoner to infer the type of the instance on the basis of these conditions. With the configuration shown in (43), the reasoner[15] can infer the instance of *pousser* as being of the more specific type $pousser\_physical_2$, as this is the only class which satisfies the condition of having agentive social objects as grammatical subject and object without any further arguments (cf. bottom of Figure 1).

Calculation of inferences. The assertion of a transitive *pousser* in combination with $subj(e, pianiste) \wedge obj(e, compositeur)$ causes the SWRL rule in Table 1 above to fire, so that the relevant inferences can be calculated and inserted into the resource. For this task we used version 7 of the Jess® rule engine[16] (see e.g. Golbreich and Imai (2004)). The result of the rule application is given below.

(44) $PUSHING(e) \wedge antagonist(e, pianiste) \wedge agonist(e, compositeur) \wedge VECTOR(v) \wedge source(v, e) \wedge LOCATION(locx) \wedge LOCATION(locy) \wedge LOCATION(locz) \wedge hasLocation(pianiste, locx) \wedge hasLocation(compositeur, locy) \wedge hasLocation(z, locz) \wedge hasStartingPoint(v, locx) \wedge hasLocation(v, locy) \wedge LINE(l1) \wedge LINE(l2) \wedge$

[15] We have used version 1.5.1 of the Pellet OWL DL reasoner (Sirin et al., 2007).
[16] `http://www.jessrules.com/`

$hasPoint(l1, locy) \wedge hasPoint(l1, locz) \wedge parallel(l1, l2) \wedge hasDirection(v, l2) \wedge hasMagnitude(v, 1)$

6 Conclusion

In this paper, we have presented an approach to modelling polysemous verbs, using standard formalisms such as OWL (Bechhofer et al., 2004) and SWRL (Horrocks et al., 2004). We have shown how the disambiguation of these verbs and their arguments can be performed in this model, and how inferences can be calculated and inserted into a representation that is capable of being interpreted by tools developed for the Semantic Web, such as the ontology editor Protégé (Knublauch et al., 2004).

The approach we propose has a number of advantages. One of these is that a very fine-grained distinction of senses based on contextual features enables accurate annotation of particular senses, and with it the calculation of inferences allowed by the respective sense. In addition to this, our approach combines an implementation of formal semantics with up-to-date technology for semantic processing and is therefore more formalised and more detailed than existing lexical semantic resources. The major drawback of our approach is, of course, the large amount of manual work required for the in-depth lexical-semantic analysis.

Although the system is – due to lack of broad coverage – not yet in a state of being applied to sophisticated reasoning tasks such as the RTE challenge (Recognising Textual Entailment; Dagan et al. (2005)), the inclusion of the contained knowledge into existing systems designed for such tasks seems very promising nonetheless. The RTE challenge consists in determining, given two text fragments, whether one text fragment is entailed by the other. In our examples of *pousser*, it is necessary to encode e.g. whether movement of the theme *boîte* in *"J'ai poussé la boîte"* is entailed or not, or whether the hypothesis *"A bomb has been placed in the basement"* can be inferred from a sentence like *"La lettre a poussé ma voisine à poser une bombe dans la cave"*. This shows that a high level of detail in the formal semantic description is a definite asset, and represents an important step beyond the information contained in existing lexical semantic resources.

References

Adler, S. and Asnès, M. (2005). Les compléments de degré en jusqu'à. *Travaux de linguistique*, 49:131–157.

Asher, N. and Lascarides, A. (1995). Metaphor in discourse. In *Proceedings of the AAAI Spring Symposium Series. Representation and Acquisition of Lexical Knowledge: Polysemy, Ambiguity and Generativity*, pages 3–7, Stanford, CA.

Baader, F., Calvanese, D., McGuinness, D. L., Nardi, D., and Patel-Schneider, P. F., editors (2003). *The Description Logic Handbook: Theory, Implementation and Applications*. CUP.

Baker, C. F., Fillmore, C. J., and Lowe, J. B. (1998). The Berkeley FrameNet project. In *Proceedings of the joint COLING/ACL 1998*, Montreal, Canada.

Bechhofer, S., van Harmelen, F., Hendler, J., Horrocks, I., McGuinness, D. L., eider, P. F. P.-S., and Stein, L. A. (2004). OWL Web Ontology Language Reference. W3C Recommendation. `http://www.w3.org/TR/owl-ref/`.

Berners-Lee, T., Hendler, J., and Lassila, O. (2001). The Semantic Web: a new form of Web content that is meaningful to computers will unleash a revolution of new possibiliti es. *Scientific American*, 284(5):34–43.

Bhatt, R. (1999). *Covert Modality in Non-Finite Contexts*. PhD thesis, University of Pennsylvania.

Boullier, P. and Sagot, B. (2005). Efficient and robust LFG parsing: SxLFG. In *Proceedings of IWPT '05*, Vancouver, BC.

Dagan, I., Glickman, O., and Magnini, B. (2005). The PASCAL Recognising Textual Entailment Challenge. In *Proceedings of the PASCAL Challenges Workshop on Recognising Textual Entailment*, Southampton, UK.

Fellbaum, C., editor (1998). *WordNet – An Electronic Lexical Database*. MIT Press, Cambridge, MA.

Franconi, E. (2003). Natural Language Processing. In Baader, F., Calvanese, D., McGuinness, D. L., Nardi, D., and Patel-Schneider, P. F., editors, *The Description Logic Handbook: Theory, Implementation and Applications*, pages 460–471. CUP.

Gangemi, A., Guarino, N., Masolo, C., and Oltramari, A. (2003a). Sweetening WordNet with Dolce. *AI Magazine*, 24(3):13–24.

Gangemi, A., Navigli, R., and Velardi, P. (2003b). The OntoWordNet Project: extension and axiomatization of conceptual relations in WordNet. In *Proceedings of ODBASE*, Catania, Italy. Springer.

Golbreich, C. and Imai, A. (2004). Combining SWRL rules and OWL ontologies with Protégé OWL Plugin, Jess, and Racer. In *Proceedings of the 7th Protégé Conference*, Bethesda, MD.

Guarino, N. and Welty, C. (2002). Evaluating Ontological Decisions with OntoClean. *Communications of the ACM*, 45(2):61–65.

Hacquard, V. (2006). *Aspects of Modality*. PhD thesis, MIT.

Horrocks, I., Patel-Schneider, P. F., Boley, H., Tabet, S., Grosof, B., and Dean, M. (2004). SWRL: A Semantic Web Rule Language Combining OWL and RuleML. W3C Member Submission. `http://www.w3.org/Submission/SWRL/`.

Jackendoff, R. S. (1990). *Semantic structures*. MIT Press, Cambridge, MA.

Karttunen, L. (1971). Implicative verbs. *Language*, 47:340–358.

Kipper-Schuler, K. (2006). *VerbNet: A broad-coverage, comprehensive verb lexicon*. PhD thesis, Computer and Information Science Dept., University of Pennsylvania, Philadelphia, PA.

Knublauch, H., Musen, M. A., and Rector, A. L. (2004). Editing description logic ontologies with the Protégé OWL plugin. In *Proceedings of DL 2004*, Whistler, BC.

Mari, A. and Martin, F. (2007). Tense, abilities and actuality entailment. In Aloni, M., Dekker, P., and Roelofsen, F., editors, *Proceedings of the XVI Amsterdam Colloquium*, pages 151–156, Universiteit Amsterdam.

Martin, F. (2008). Forceful verbs in Spatial, Psychological and Conceptual Domains. Semantic Analysis of *push*-verbs in French. Manuscript, University of Stuttgart.

Niles, I. and Pease, A. (2001). Towards a Standard Upper Ontology. In *Proceedings of the 2nd International Conference on Formal Ontology in Information Systems (FOIS 2001)*, Ogunquit, ME.

Niles, I. and Pease, A. (2003). Linking Lexicons and Ontologies: Mapping WordNet to the Suggested Upper Merged Ontology. In *Proceedings of the 2003 International Conference on Information and Knowledge Engineering (IKE '03)*, Las Vegas, NV.

O'Connor, M., Knublauch, H., Tu, S., Grosof, B., Dean, M., Grosso, W., and Musen, M. (2005). Supporting Rule System Interoperability on the Semantic Web with SWRL. In *Proceedings of the 4th International Semantic Web Conference*, Galway, Ireland.

Sirin, E., Parsia, B., Grau, B. C., Kalyanpur, A., and Katz, Y. (2007). Pellet: A practical OWL-DL reasoner. *Journal of Web Semantics*, 5(2).

Sonntag, D., Engel, R., Herzog, G., Pfalzgraf, A., Pfleger, N., Romanelli, M., and Reithinger, N. (2007). SmartWeb Handheld – Multimodal Interaction with Ontological Knowledge Bases and Semantic Web Services. In Huang, T. S., Nijholt, A., Pantic, M., and Pentland, A., editors, *Artificial Intelligence for Human Computing*, volume 4451 of *Lecture Notes in Artificial Intelligence*, pages 272–295. Springer, Heidelberg.

Spohr, D. (2008a). A General Methodology for Mapping EuroWordNets to the Suggested Upper Merged Ontology. In *Proceedings of the 6th Language Resources and Evaluation Conference (LREC 2008)*, Marrakech, Morocco.

Spohr, D. (2008b). Extraction of Selectional Preferences for French using a Mapping from EuroWordNet to the Suggested Upper Merged Ontolog y. In *Proceedings of the 4th Global WordNet Conference*, Szeged, Hungary.

Stein, A. (2007). Motion events in concept hierarchies: Identity criteria and french examples. In Schalley, A. and Zaefferer, D., editors, *Ontolinguistics. How Ontological Status Shapes the Linguistic Coding of Concepts*, pages 379–394. Mouton de Gruyter, Berlin.

Talmy, L. (1985). Lexicalization patterns: semantic structure in lexical form. In Shopen, T., editor, *Language typology and syntactic description III*, pages 51–149. Cambridge University Press, New York, NY.

Vandeloise, C. (1991). *Spatial Prepositions: a Case Study from French*. University of Chicago Press, Chicago.

Vossen, P., editor (1998). *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*. Kluwer Academic Publishers.

Zwarts, J. (2007). Forceful prepositions. In Vyvyan, E. and Chilton, P., editors, *Language, Cognition and Space: The State of the Art and New Directions*. Equinox Publishing, London. To appear.

Author Index

Lucie Barque
Université Lille 3, France
lucie.barque@gmail.com

Fran cois-Regis Chaumartin
Proxem, France
webmaster@proxem.com

Lars-Trap Jensen
Det Danske Sprog- og Litteraturselskab, DSL
ltj@dsl.dk

Ernesto William de Luca
Distributed Artificial Intelligence Laboratory, TU Berlin
ernesto.deluca@dai-labor.de

Fabienne Martin
Institut für romanistische Linguistik, Universität Stuttgart
fabienne.martin@ling.uni-stuttgart.de

Sanni Nimb
Det Danske Sprog- og Litteraturselskab, DSL
sn@dsl.dk

Andreas Nürnberger
Fakultät für Informatik, Otto-von Guericke Universität Magdeburg
andreas.nuernberger@ovgu.de

Kiril Simov
Institute for Parallel Processing, Bulgarian Academy of Sciences, Sofia, Bulgaria
simov@bultreebank.org

Dennis Spohr
Institut für romanistische Linguistik, Universität Stuttgart
dennis.spohr@ling.uni-stuttgart.de

Achim Stein
Institut für romanistische Linguistik, Universität Stuttgart
achim.stein@ling.uni-stuttgart.de

Nofiza Vokhidova
Universität Mannheim
n.vokhidova@yahoo.de

View publication stats